KAIST

Advanced Machine Intelligence Lab.

Multimodal Generation Machine Perception

- Beyond Scale -

Explicit Priors for Semantic & Physical Understanding in Audio-Visual Learning

---

Tae-Hyun Oh

Associate Professor
School of Computing
KAIST

Prof. Tae-Hyun Oh @ KAIST

http://ami.kaist.ac.kr

Advanced Machine Intelligence Lab.

Advanced Machine Intelligence Lab.



PI

Tae-Hyun Oh

Associate Professor

Ph.D. students	17	Postdoc	5
M.S. students	11	Admin staff	2







Prof. Tae-Hyun Oh @ KAIST

http://ami.kaist.ac.kr



Tae-Hyun Oh

Associate Professor

KAIST

Experience

Ph.D. degree, KAIST	12-17
Assoc. prof., School of Computing, KAIST	25-Now
Assist./Assoc. prof., Dept. EE, POSTECH,	20-25
Postdoc researcher, Facebook AI Research (FAIR),	19-20
Postdoc associate, MIT CSAIL,	17-19
Research interns, MSR Redmond & MSR Asia,	14-15, 16

Research Interest

Vision-centric Multi-modal Perception
Physics-aware AI for Embodied Agents
Synthetic data generation

Award

BMVC Best Poster Award ‘24
Microsoft Fellowship Award ‘15
Samsung Humantech Paper Award (Gold) ‘15
KAIST EE Kim Chung Ki Award ‘15

Academic Service

Area chair - CVPR, ICCV, NeurIPS, ICML, ICLR, ACCV
Senior PC - AAAI, IJCAI, SIGGRAPH Asia
Associate editor - TPAMI, IJCV, TMLR

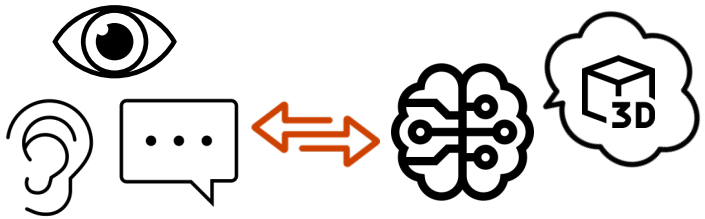
Advanced Machine Intelligence Lab.

Advanced Machine Intelligence Lab.
PI: Tae-Hyun Oh

Multi-modal AI

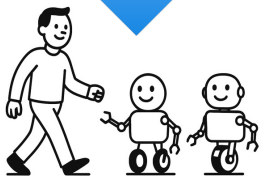
- Multi-modal Perception
- Spatial & Physics-aware AI

Research Goal



Multi-modal Perception

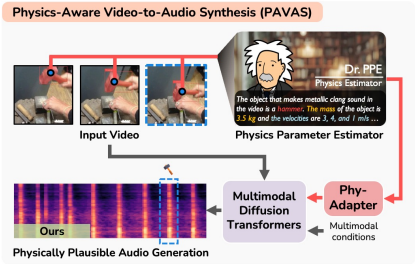
Spatial & Physics-aware AI



Multi-modal Embodied-AI with Spatial and Physics Reasoning

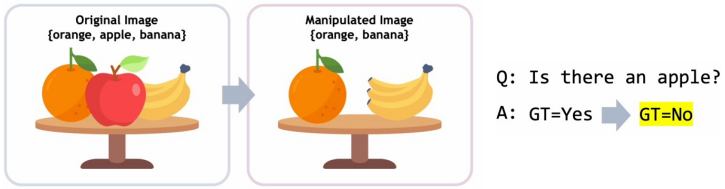
Multi-modal Perception

Audio-Visual Understanding



Physics-aware video-to-audio synthesis
(CVPR 2026 Oral)

Vision-Language Understanding



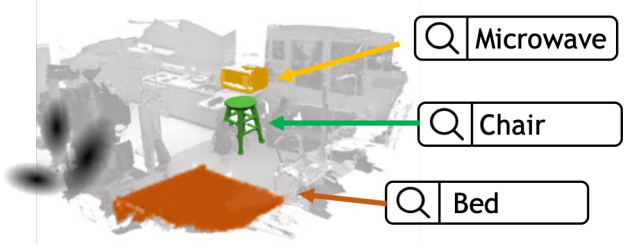
Evaluate hallucination in VLMs[2]
(ECCV 2024)

[1] Hyun-Bin et al. "PAVAS: Physics-Aware Video-to-Audio Synthesis." CVPR 2026

[2] Ye-Bin et al. "Beaf: Observing before-after changes to evaluate hallucination in vision-language models." ECCV 2024

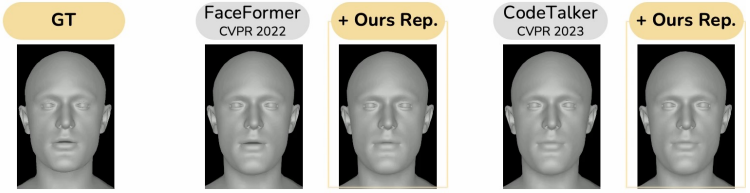
Spatial & Physics-aware AI

3D Scene Understanding



Language-embedded 3D scene[3]
(CVPR 2025 Highlight)

3D Human Understanding



3D talking head representation[4]
(CVPR 2025 Highlight)

[3] Jun-Seong et al. "Dr. Splat: Directly Referring 3D Gaussian Splatting via Direct Language Embedding Registration." CVPR 2025

[4] Chae-Yeon et al. "Perceptually Accurate 3D Talking Head Generation: New Definitions, Speech-Mesh Representation, and Evaluation Metrics." CVPR 2025

Research Theme

Learning to visualize invisible signals

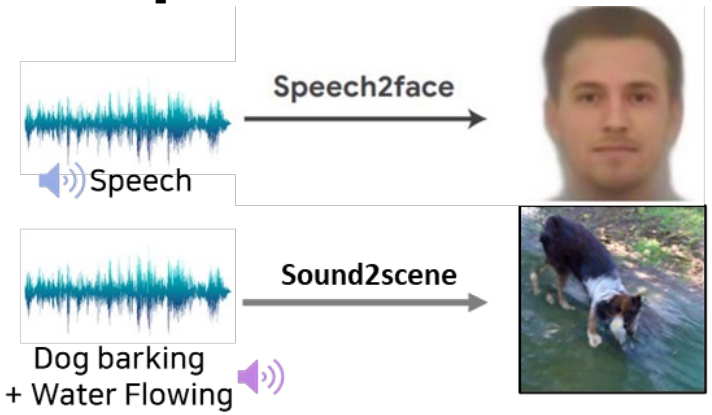
Summary

- Invisible signals: sound, text, invisibly small things, **and your mind**
- Invisible-to-visible ⇒ Cross-modal translation ⇒ 1st step toward AGI

The Sound Camera

[CVPR'19,'23] [ICASSP'23, AAAI'25]

YTN · YONHAP NEWS AGENCY · GIZMODO

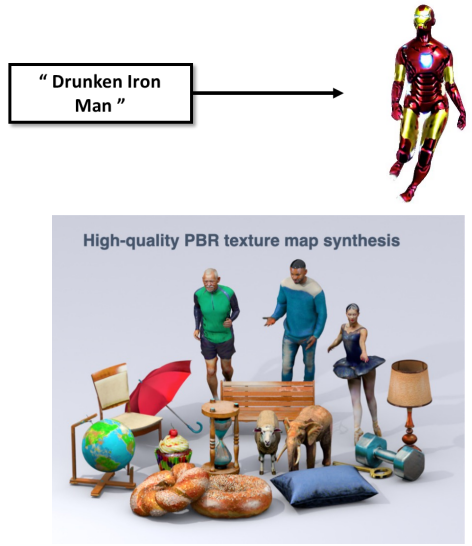


Text-to-3D/4D

[ECCV'22, CVPR'24]

Qualcomm Innovation Fellowship

전자신문 논문상 대상(과기부 장관상)



Subtle motion magnification

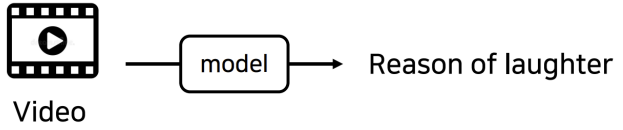
[ECCV'18 (Oral), '24, ICLR'23]

Humantech Paper Award



Theory of Mind via the lens of MLLM

Understand why & how you laugh [NAACL'24]



KAIST

Advanced Machine Intelligence Lab.

- Beyond Scale -
Explicit Priors for Semantic & Physical Understanding in Audio-Visual Learning

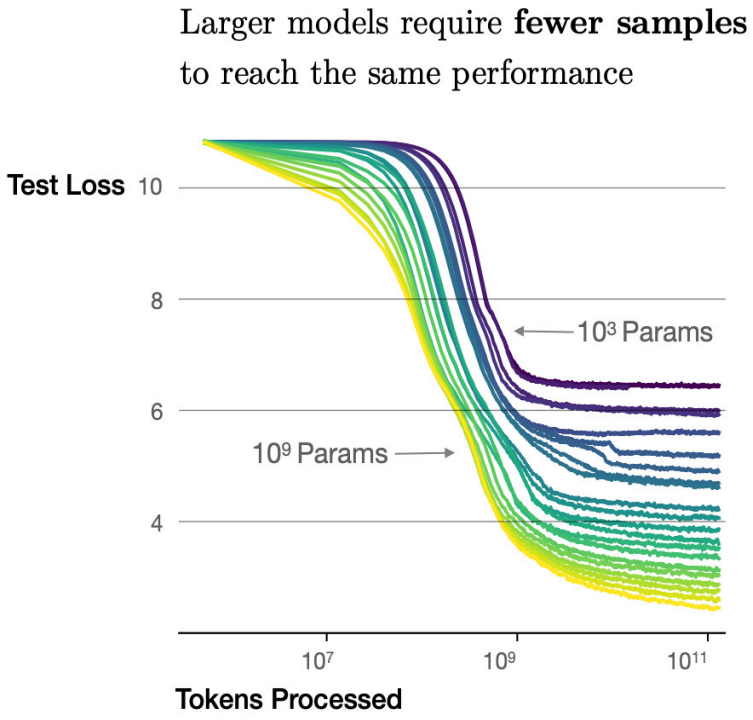
---

Tae-Hyun Oh

Associate Professor
School of Computing
KAIST

Today's Keywords – Scaling Law

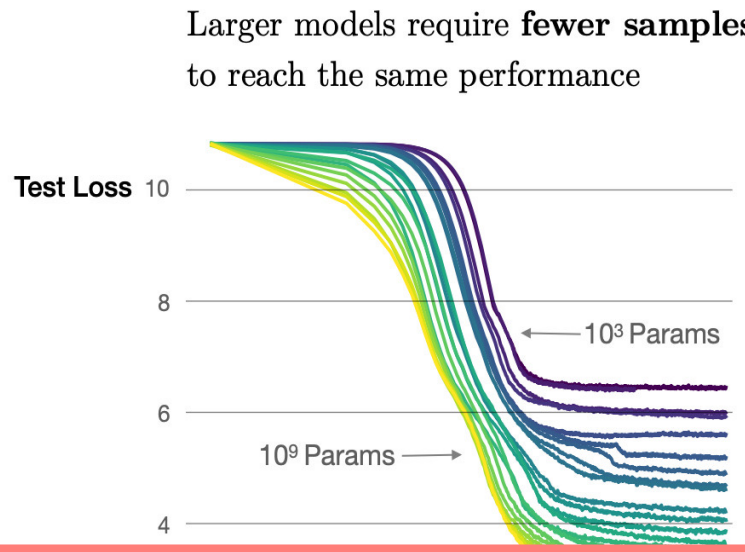
- Larger data + A larger model = You win!
- Typically done with self-supervised learning





Today's Keywords – Scaling Law

- Larger data + A larger model = You win!
- Typically done with self-supervised learning





- Scaling Laws -
Necessary, but NOT Sufficient

Today’s Keywords - Audio-visual learning (AVL)

- Multiple reasons why we are interested in AVL
  - Creative Applications in Entertainment and Media

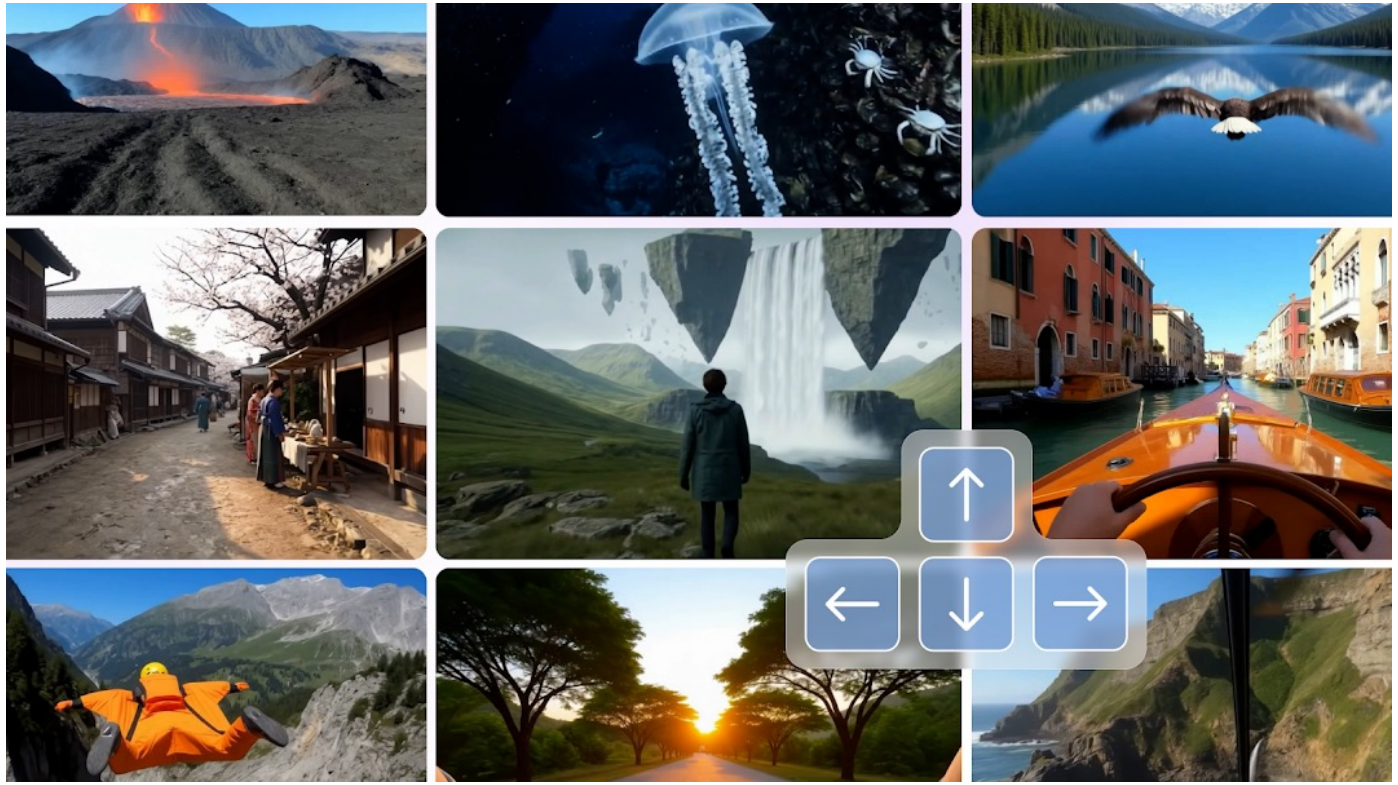




Veo3

Today’s Keywords - Audio-visual learning (AVL)

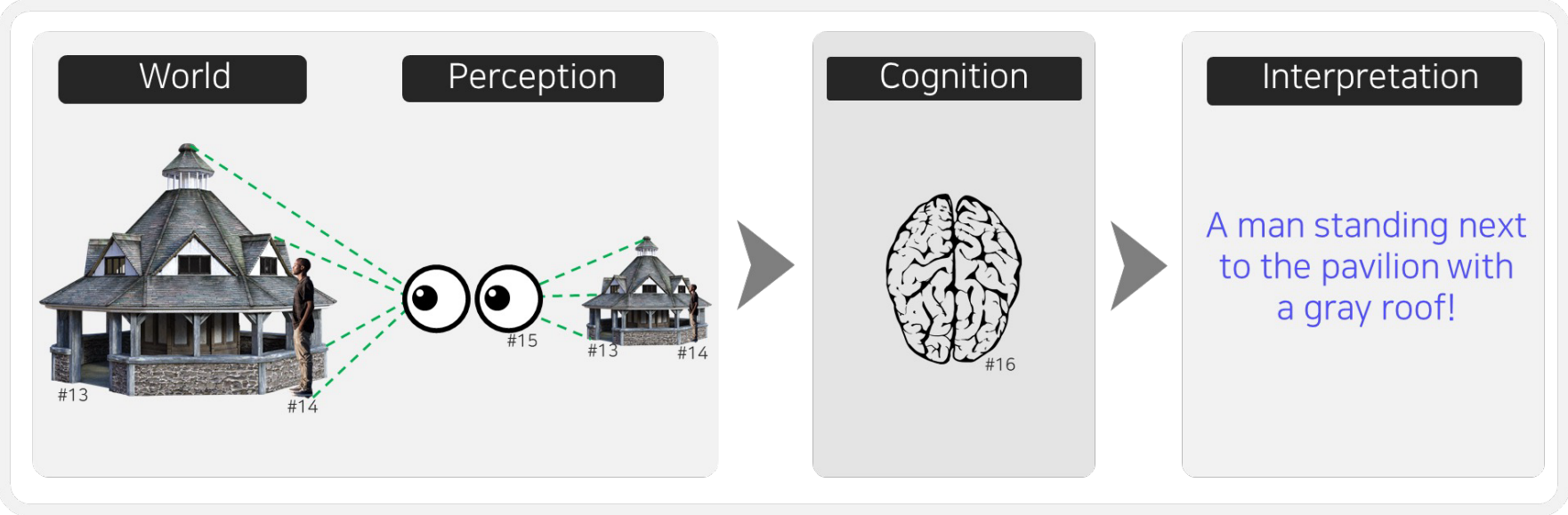
- Multiple reasons why we are interested in AVL
  - Creative Applications in Entertainment and Media
  - Toward Multimodal World Model (Genie 3 + Veo 3?)



Genie 3

Today's Keywords - Audio-visual learning (AVL)

- Multiple reasons why we are interested in AVL
  - Creative Applications in Entertainment and Media
  - Toward Multimodal World Model (Genie 3 + Veo 3?)
  - **Multimodal Machine Perception**



Computer Vision



Common Cameras

capture the visible



Common AI

understand the visible

Computer Vision



Common Cameras

capture the visible



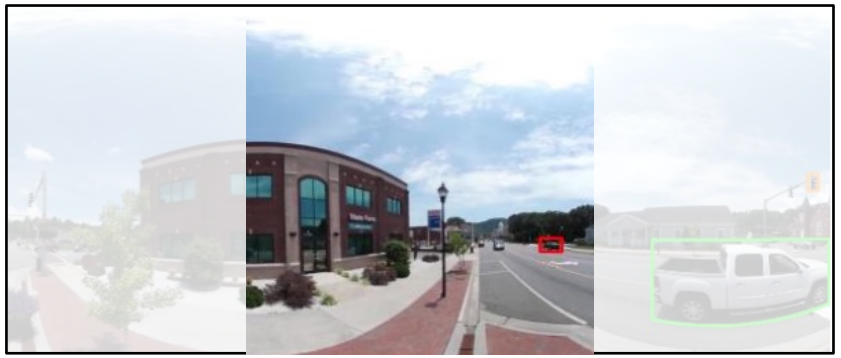
Common AI

understand the visible

Is this sufficient to understand our surrounding world?

Limitation of Computer Vision

- Cameras (also, our eyes) capture limited sights due to
  - Limited field of view
  - Occlusion



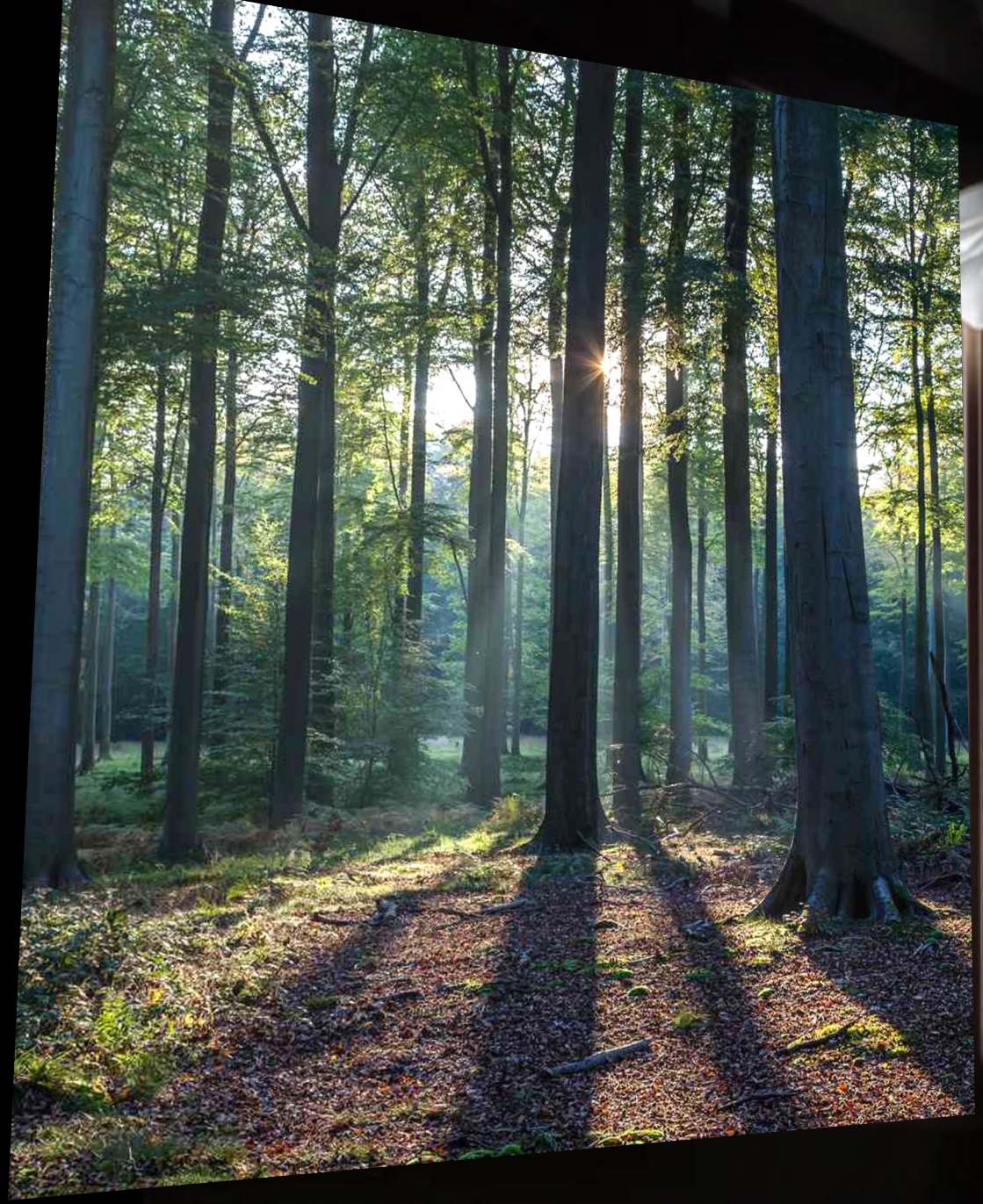
Hidden information beyond field of view



Occlusion

What we see is only a fraction of the information around us







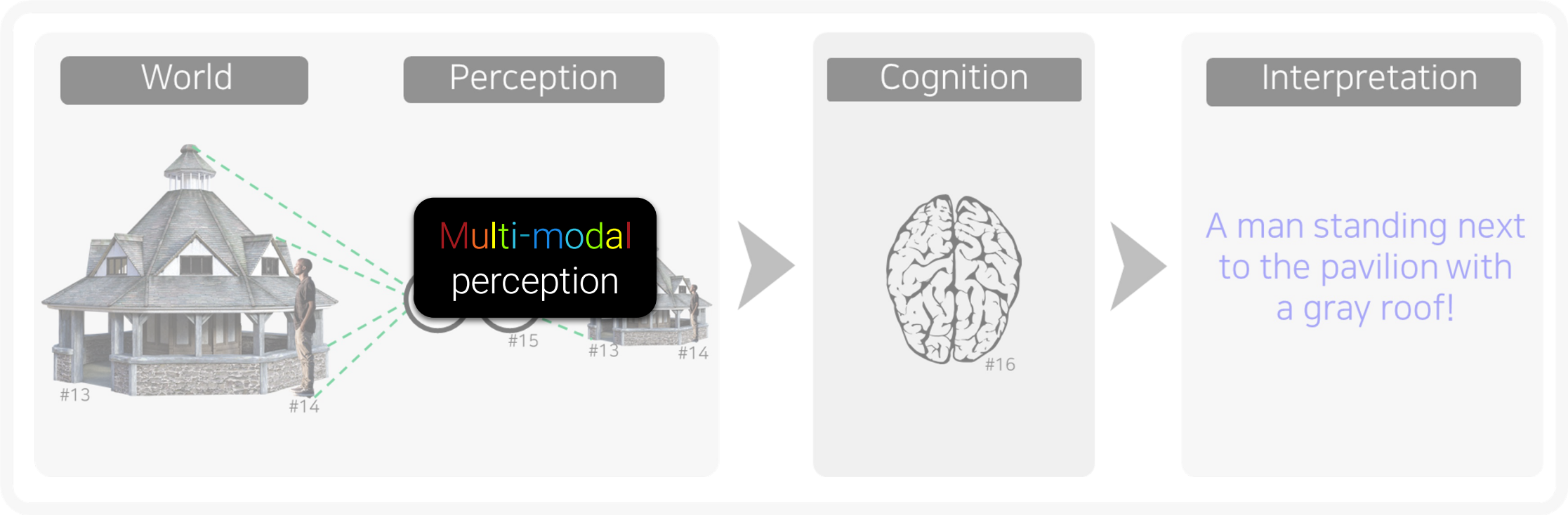
Multimodal perception system

cognition

perception

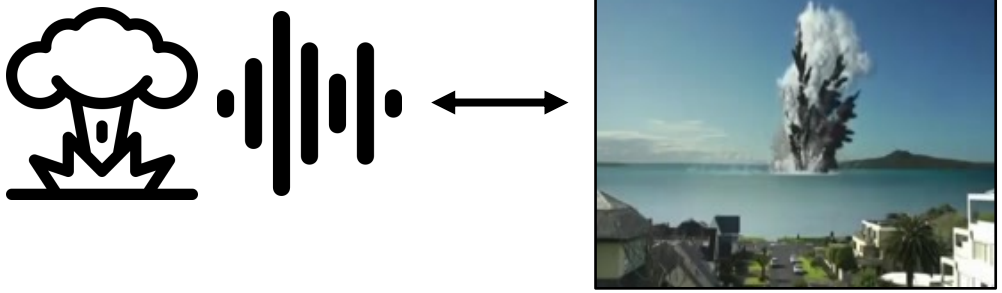
higher cognitive functions

Other signals help compensate for missing or weak modalities



Human's scene understanding process

Sound is everywhere

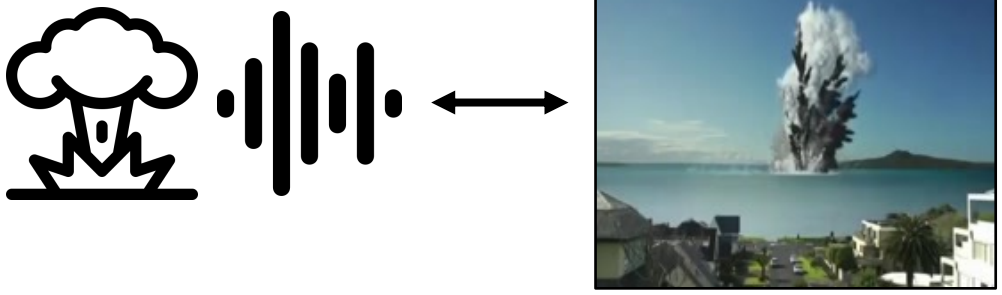


Audio event

Visual event

Natural co-occurrence

Sound is everywhere



Audio event Visual event

Natural co-occurrence



Vision + Sound

Non-line-of-sight information

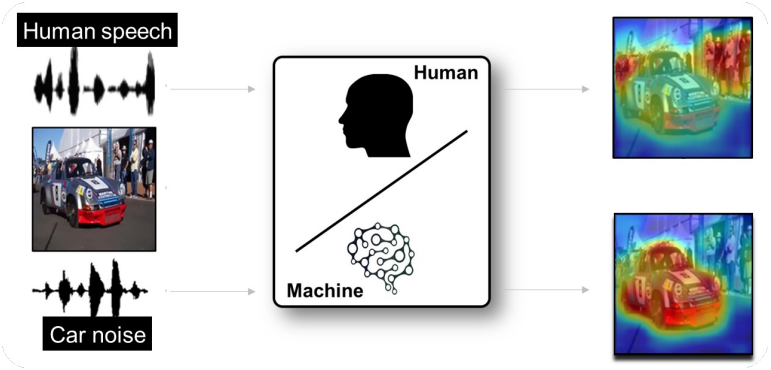
Sound conveys important event information around us


The Sound Camera

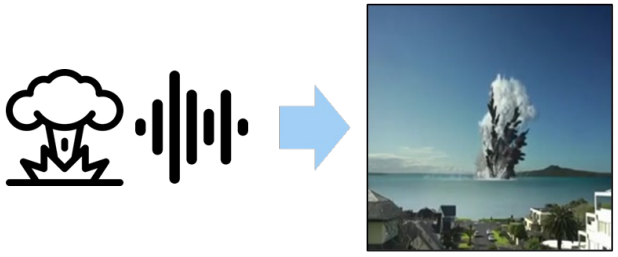
Learning Audio-Visual Perception to Sense the World

Today’s contents

Semantic understanding
[TPAMI’21, CVPR’23]



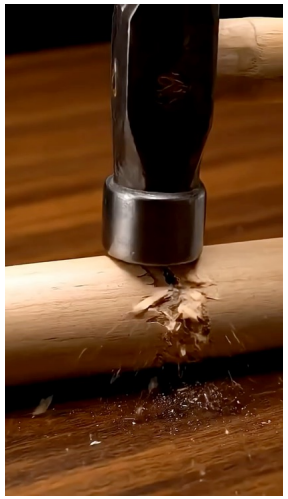
Sound source localization



Audio event Visual event

Sound-to-Visual event generation

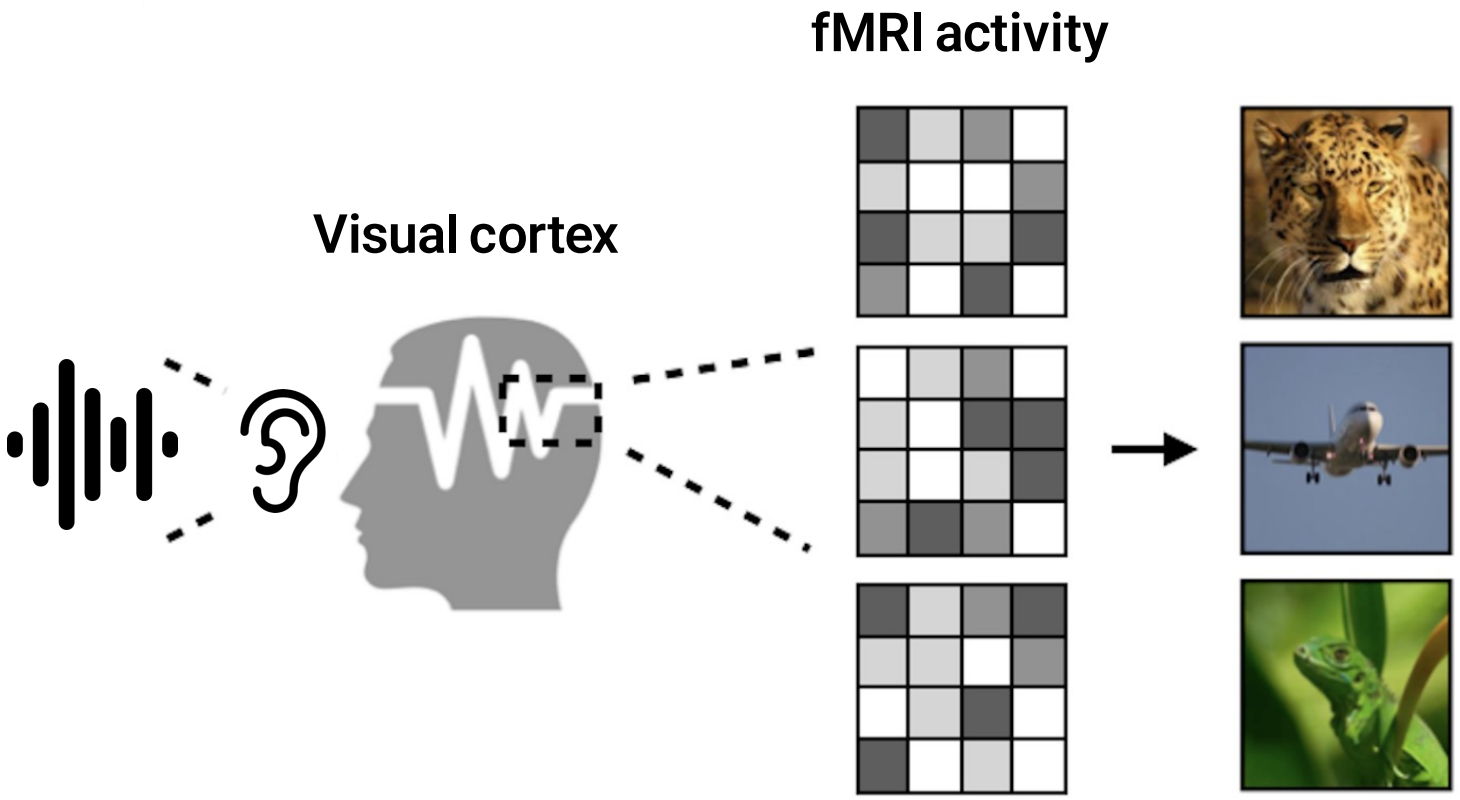
Physical understanding
[CVPR’26 Oral]



Video-to-Sound generation

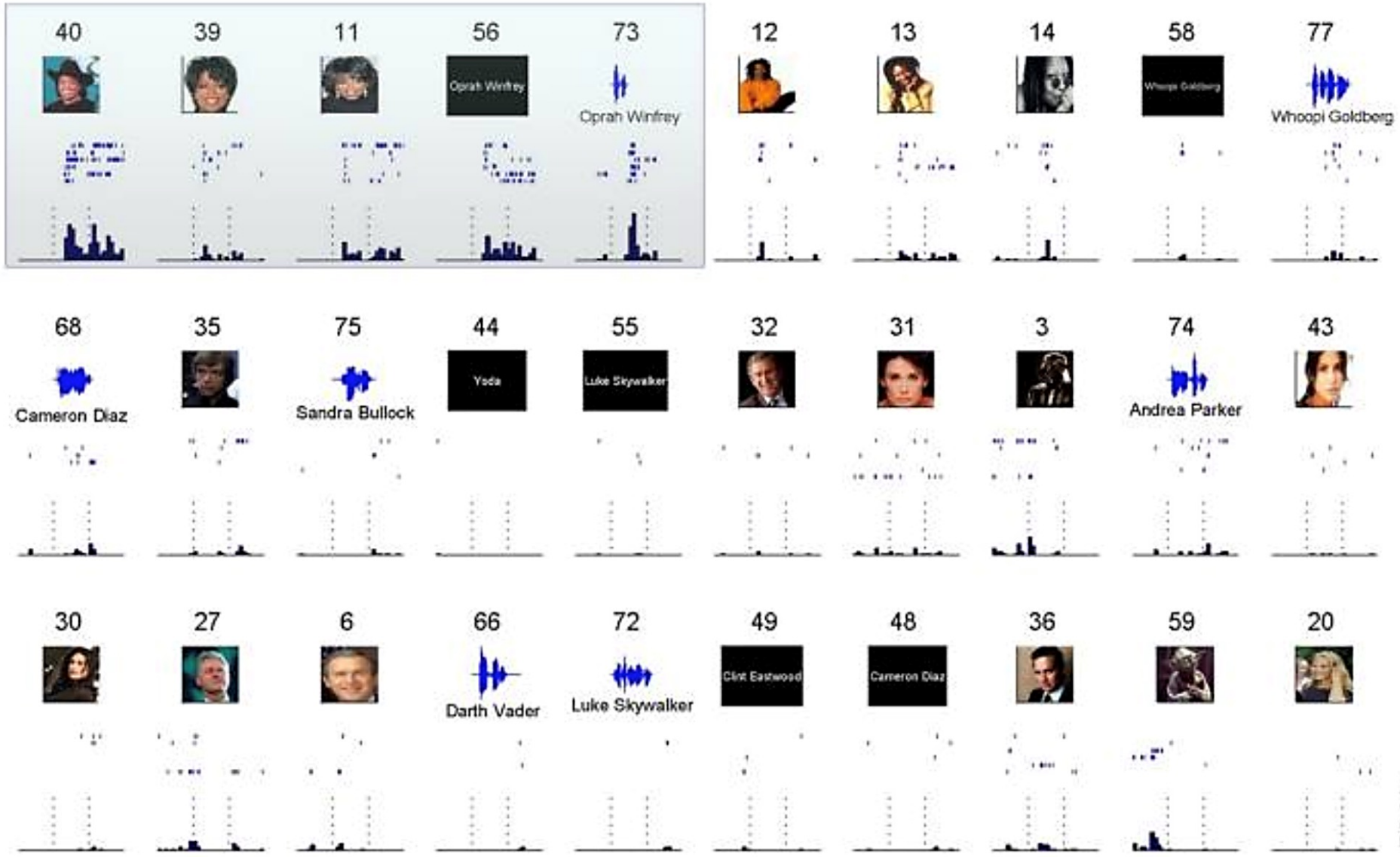
How do we translate sound

Salient Sounds Activate Human Visual Cortex Automatically [MacDonald13]



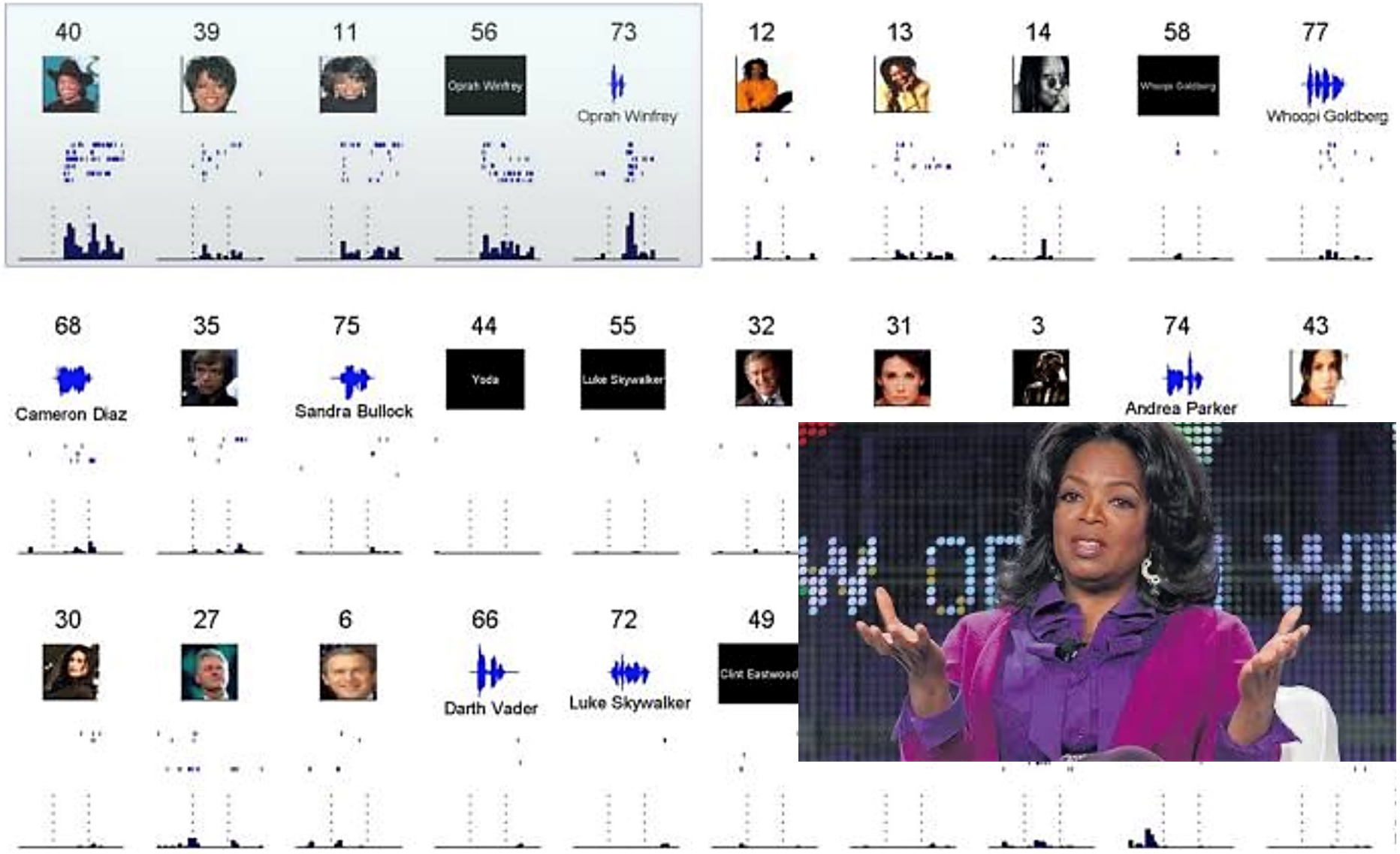
MacDonald et al. Salient Sounds Activate Human Visual Cortex Automatically, The Journal of Neuroscience, 2013

Oprah Winfrey Neuron



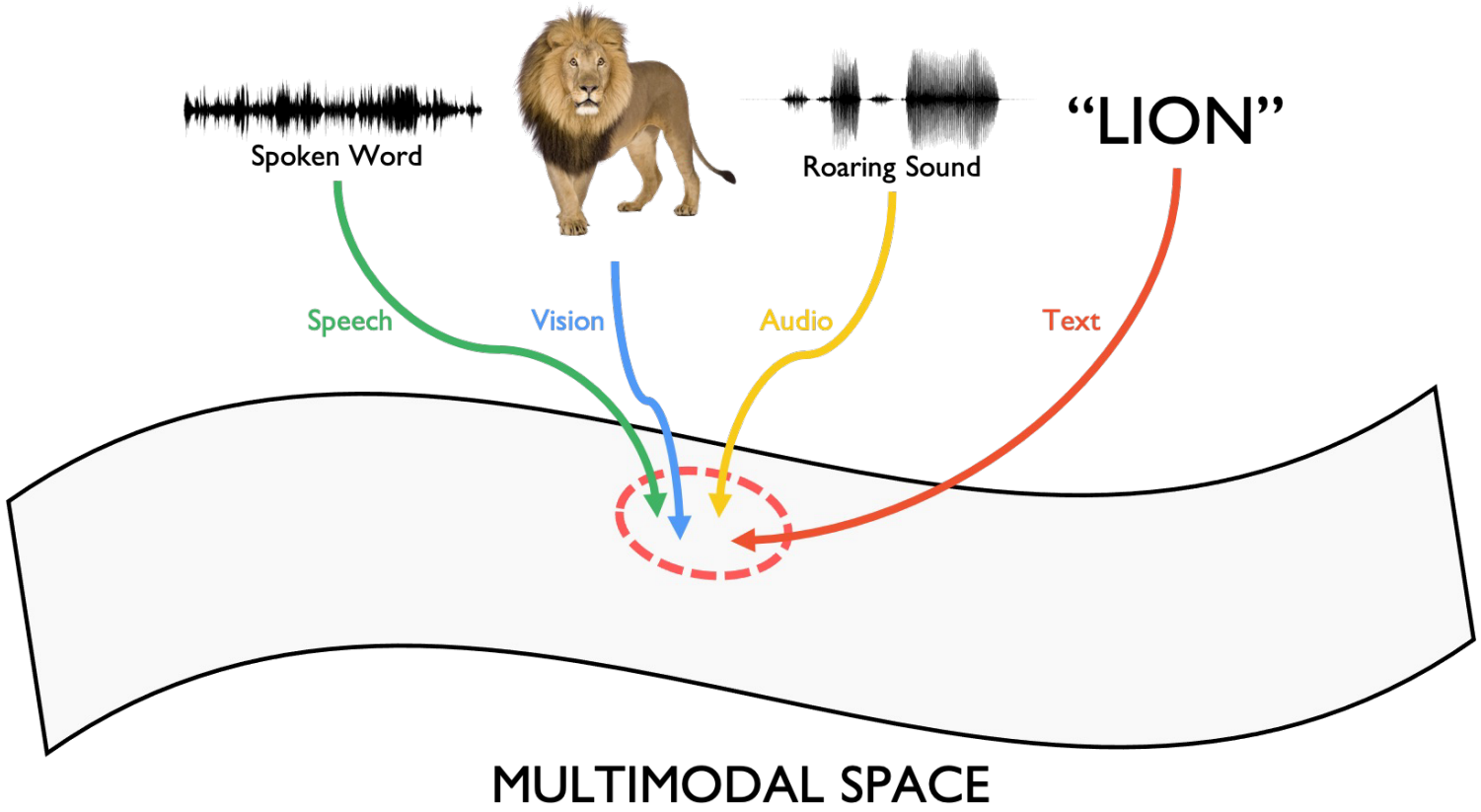
Quiroga et al. Explicit Encoding of Multimodal Percepts by Single Neurons in the Human Brain, Current Biology, 2009

Oprah Winfrey Neuron



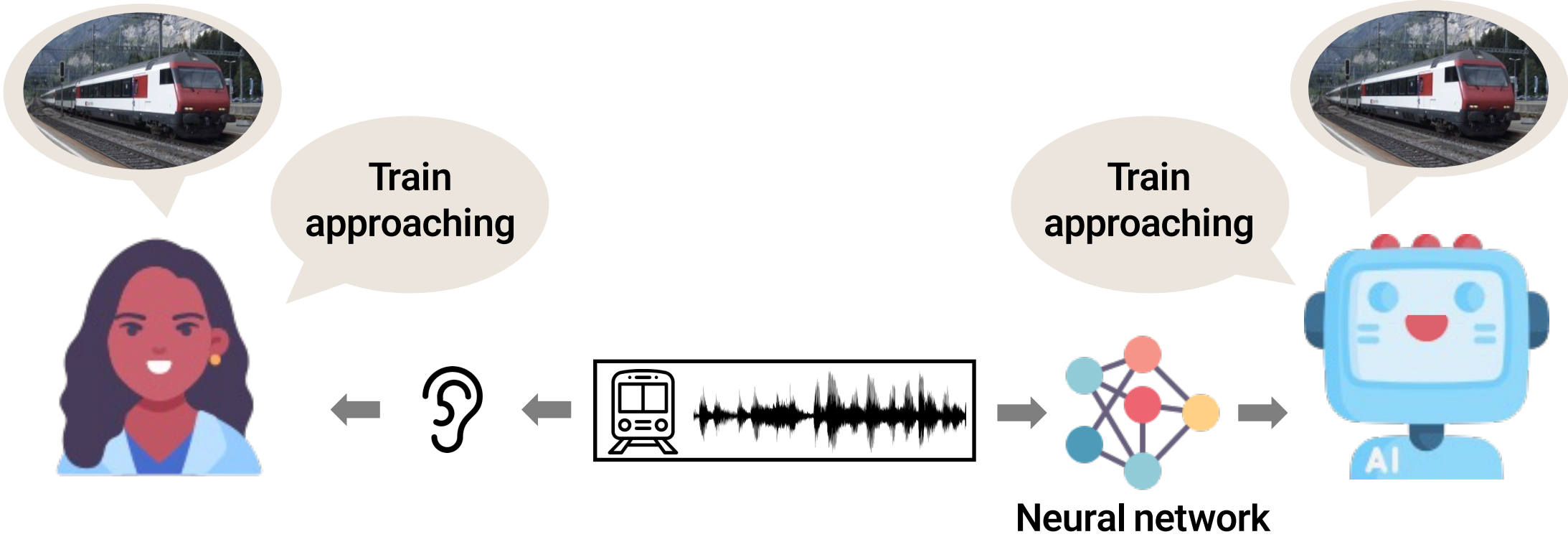
Quiroga et al. Explicit Encoding of Multimodal Percepts by Single Neurons in the Human Brain, Current Biology, 2009

Multisensory Fusion for Concept Learning

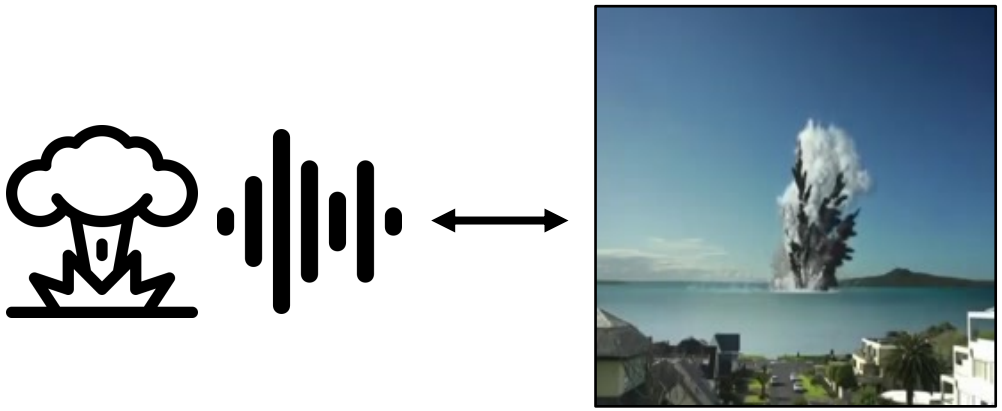


GA Calvert. Crossmodal Processing in the Human Brain: Insights from Functional Neuroimaging Studies, Cerebral Cortex, 2001
Quiroga et. al. Invariant Visual Representation by Single Neurons in the Human Brain, Nature, 2005
Quiroga et. al. Explicit Encoding of Multimodal Percepts by Single Neurons in the Human Brain, Current Biology, 2009
Choksi et. al. Multimodal Neural Networks Better Explain Multivoxel Patterns in the Hippocampus, NeurIPS Workshop, 2021
Goh et. al. Multimodal Neurons in Artificial Neural Networks, OpenAI, 2021

Build human-like machine perception



Sound is everywhere

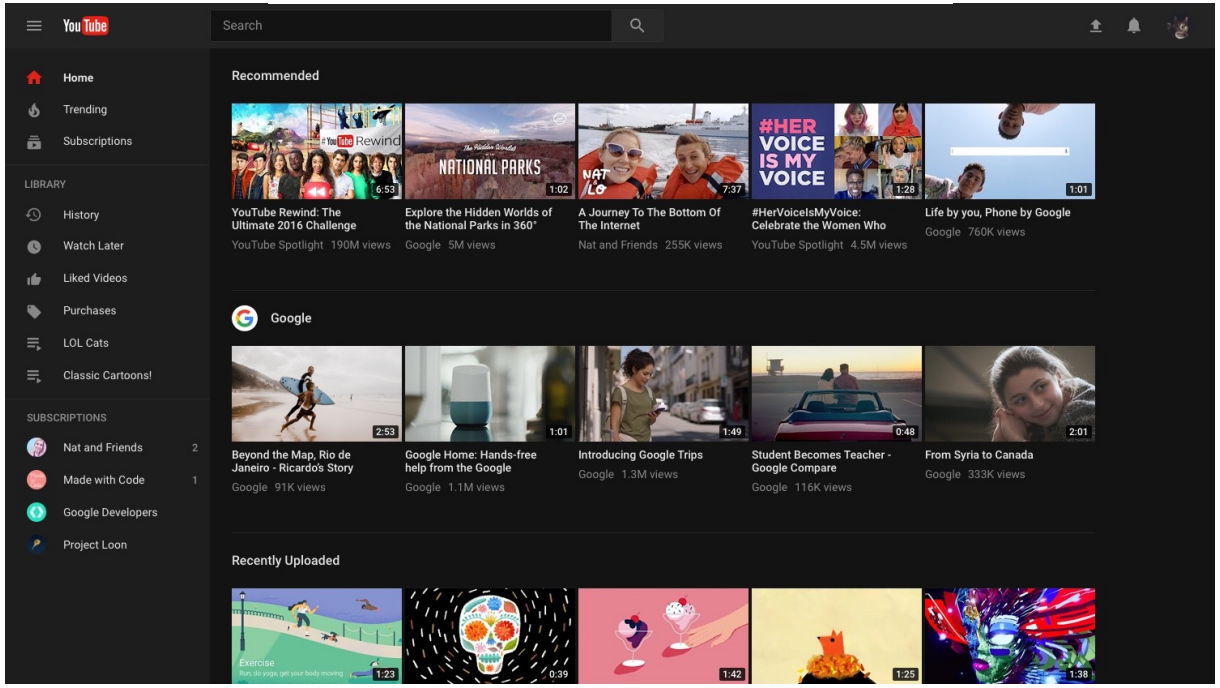


Audio event

Visual event

Natural co-occurrence (Good data source)

YouTube



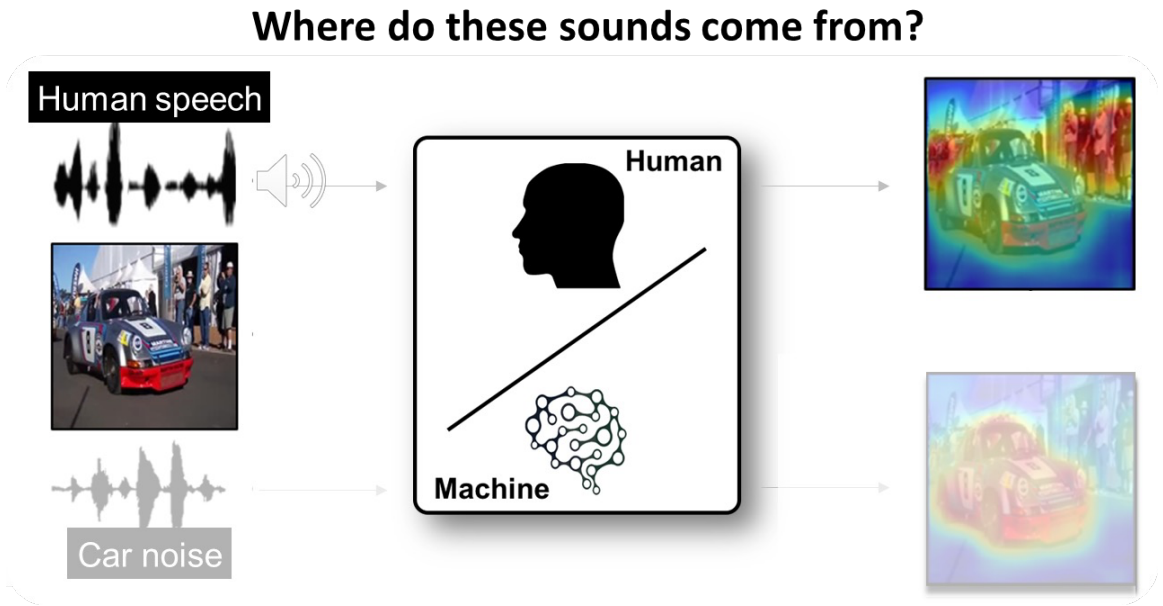
Semantic Understanding in Audio-Visual Learning

IEEE TRANSACTIONS ON
PATTERN ANALYSIS AND
MACHINE INTELLIGENCE

Learning to Localize Sound Sources in Visual Scenes

Arda SENOCAK[1] Tae-Hyun Oh[2] Junsik Kim[1] Ming-Hsuan Yang[3] In So Kweon[1]

[1]KAIST [2]MIT CSAIL [3]UC. Merced



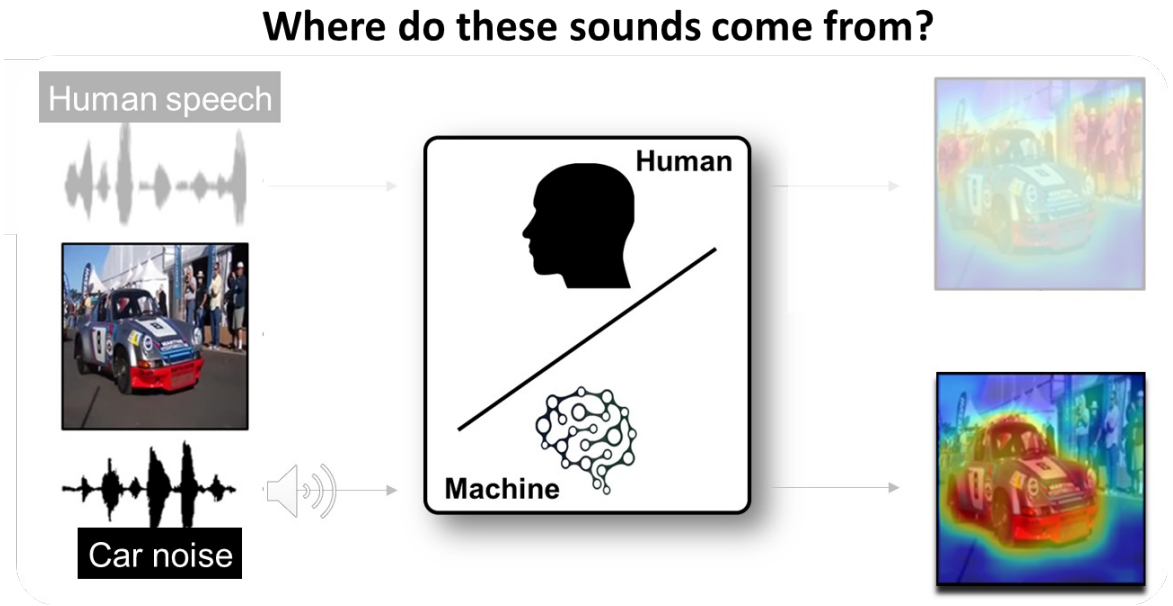
Arda et al., Learning to Localize Sound Sources in Visual Scenes: Analysis and Applications, TPAMI 2021

IEEE TRANSACTIONS ON
PATTERN ANALYSIS AND
MACHINE INTELLIGENCE

Learning to Localize Sound Sources in Visual Scenes

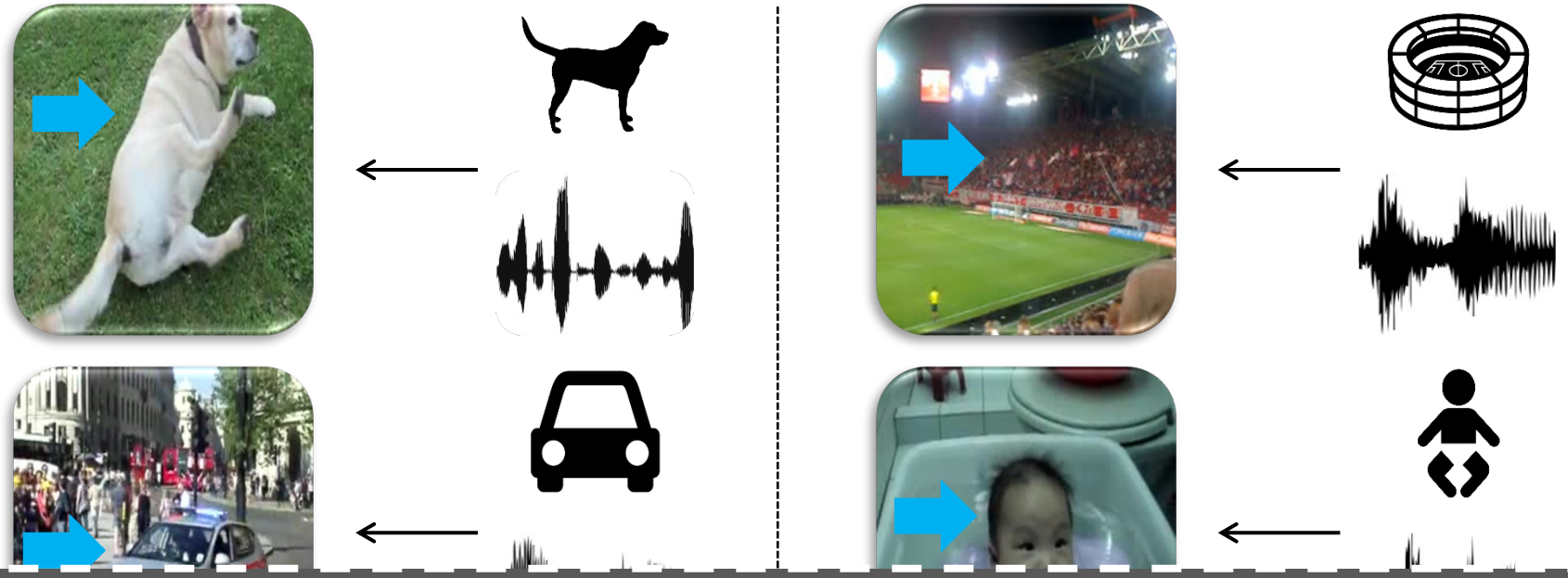
Arda SENOCAK[1] Tae-Hyun Oh[2] Junsik Kim[1] Ming-Hsuan Yang[3] In So Kweon[1]

[1]KAIST [2]MIT CSAIL [3]UC. Merced



Arda et al., Learning to Localize Sound Sources in Visual Scenes: Analysis and Applications, TPAMI 2021

Humans understand what's producing sound

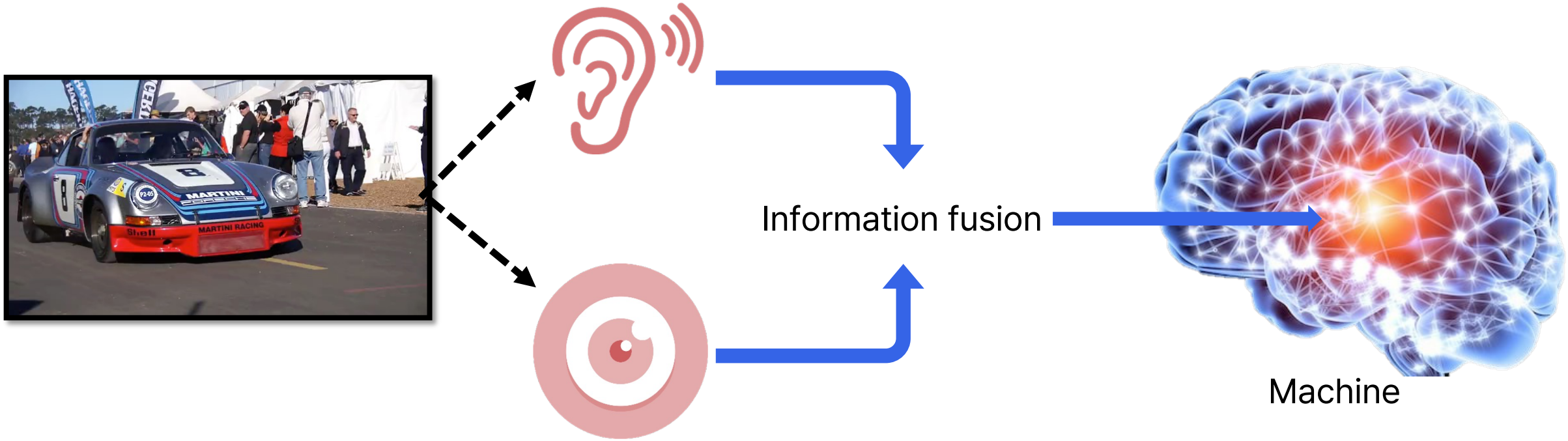


Can a computer **learn to localize sound sources just by watching** videos with their coupled audio?

Learns “co-occurrence” by watching and hearing

- Audio-visual correspondence [Arandjelović and Zisserman, ICCV17]

Is this a learnable problem by co-occurrence (self-supervision) alone?



Unified architecture

- Can be used for testing

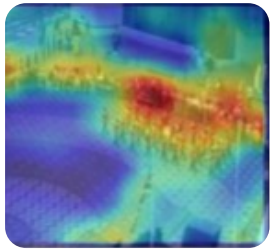
  { **Self** (Un)-supervision / **Full** supervision / **Semi-**supervision }

Image



Audio



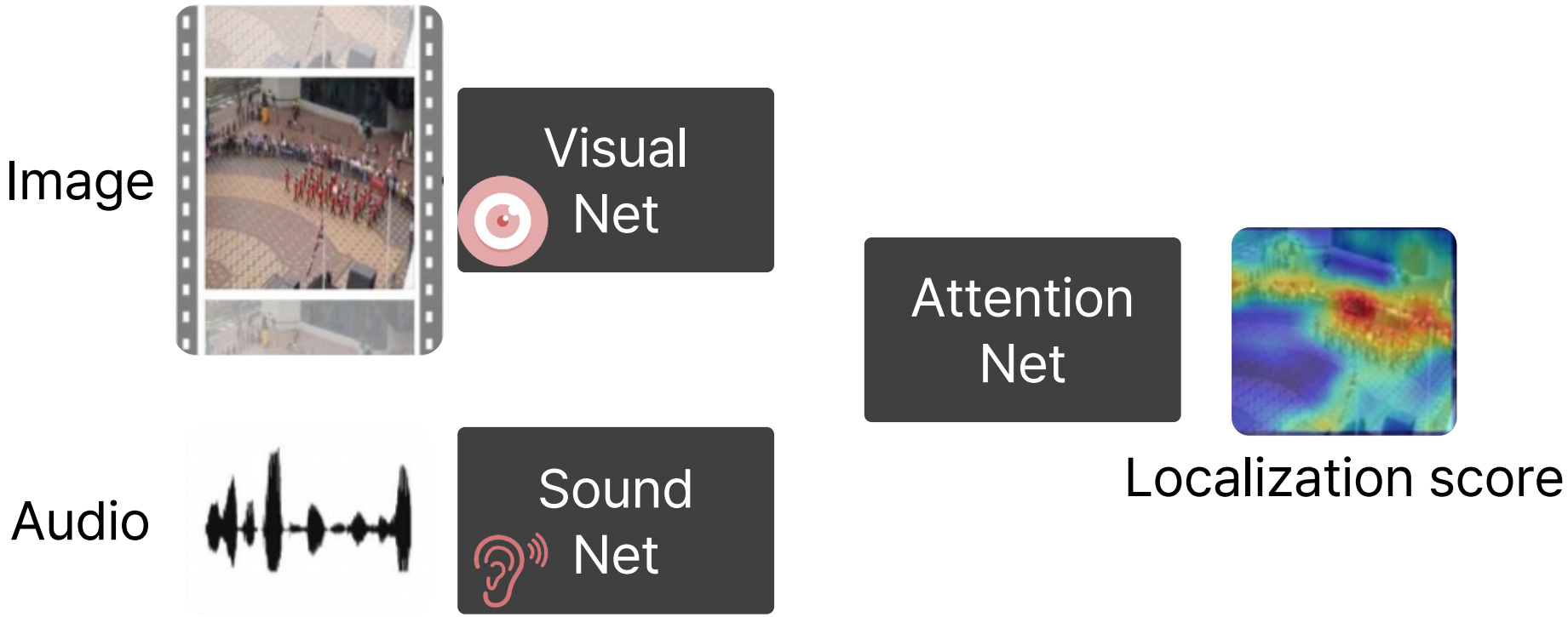


Localization score

Unified architecture

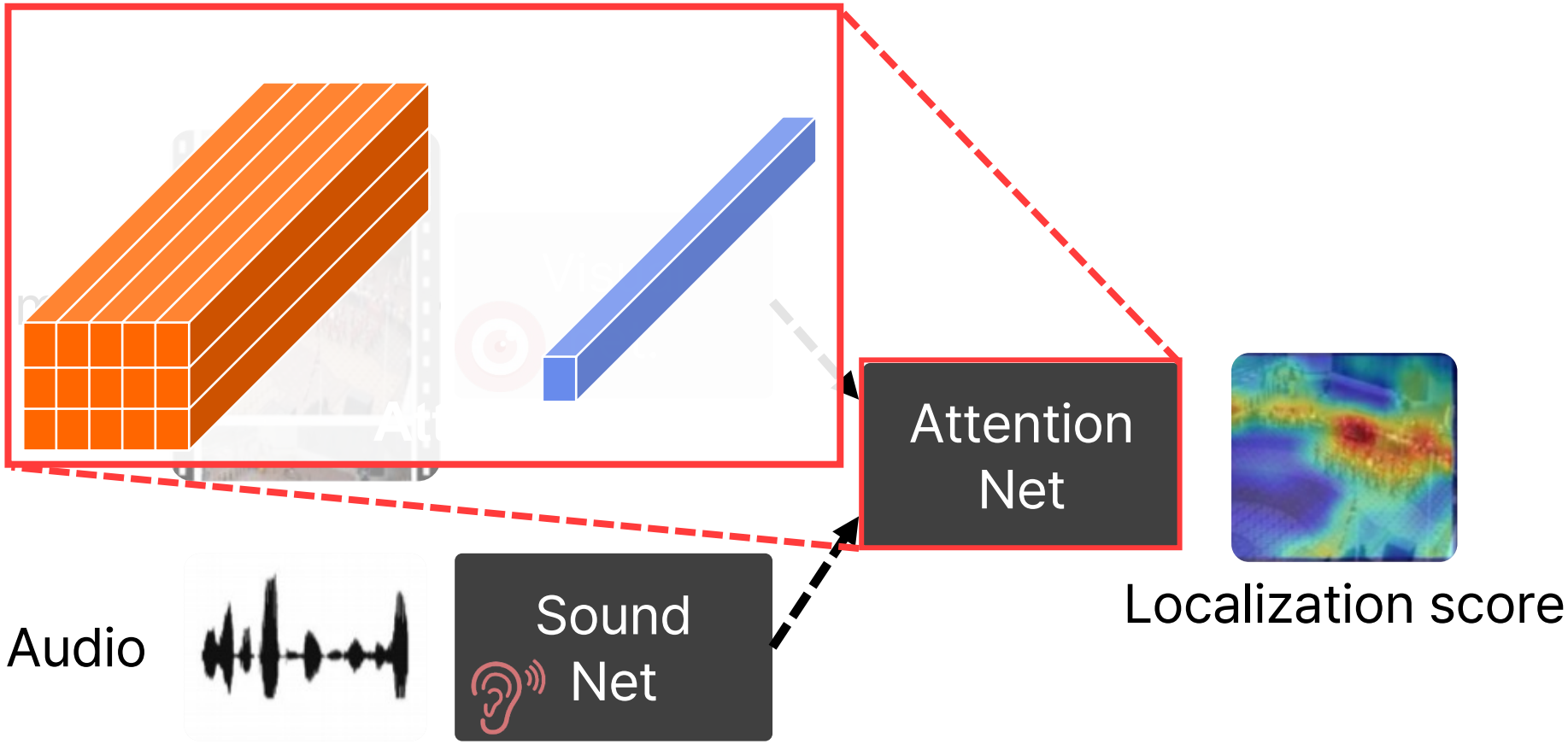
- Motivated by human cognition structure

  [Shelton et al. The influence of vision on the absolute identification of sound-source position. Perception & Psychophysics, 1980]



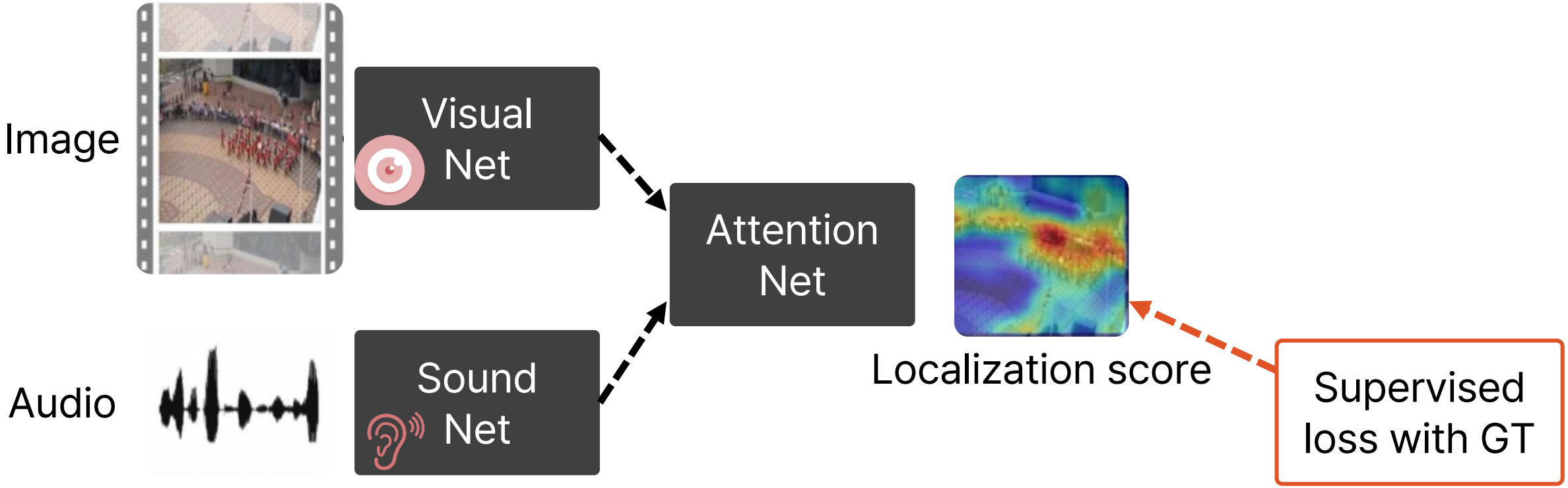
Unified architecture

- Attention layer measures correlation at each spatial location



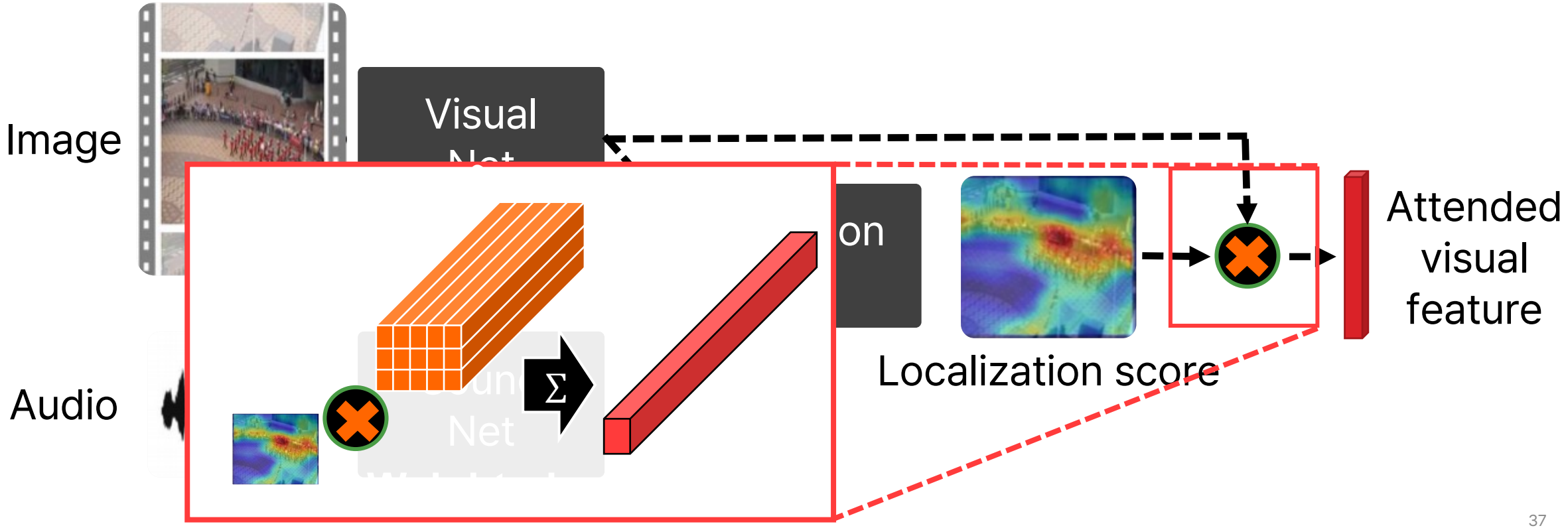
Unified architecture

- Supervised learning case



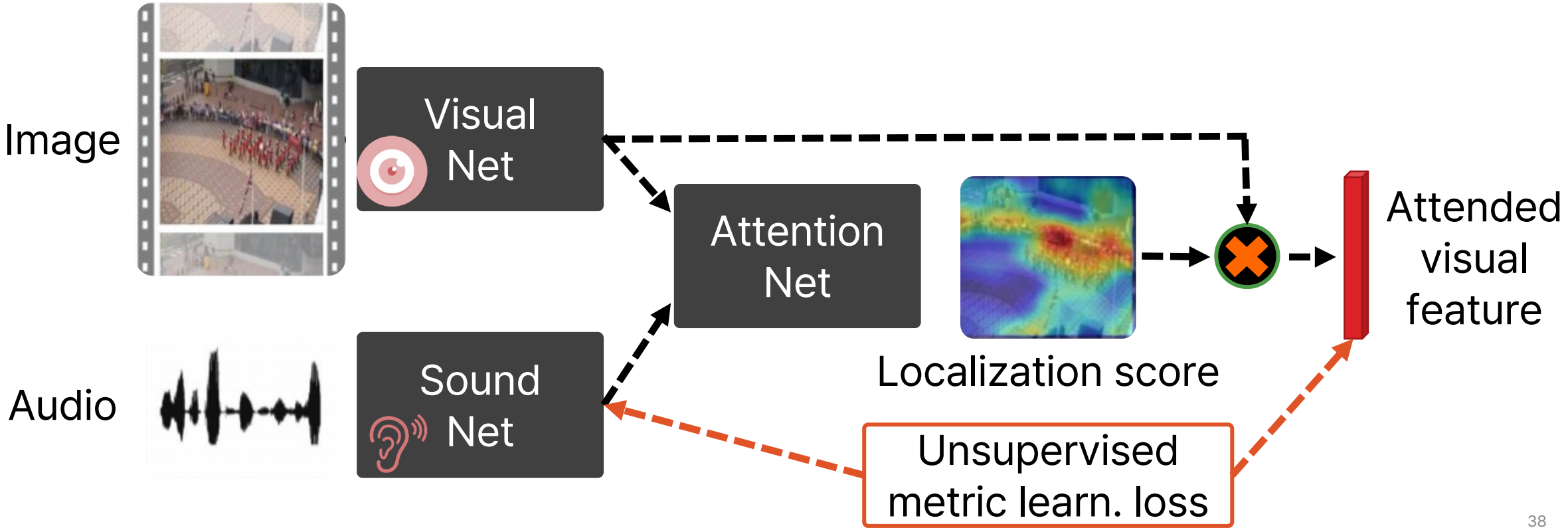
Unified architecture

- Un-supervised learning case
  - Attention based pooling extracts
    a feature vector corresponding to localized region



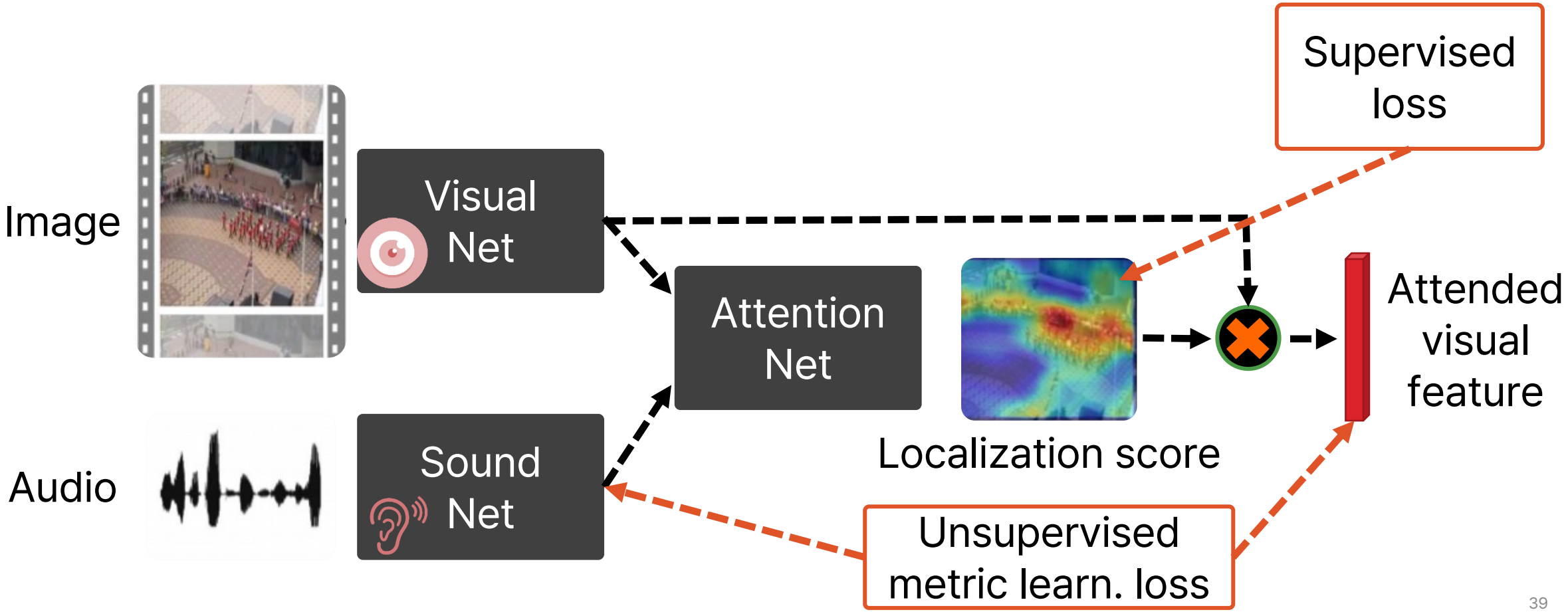
Unified architecture

- Un-supervised learning case



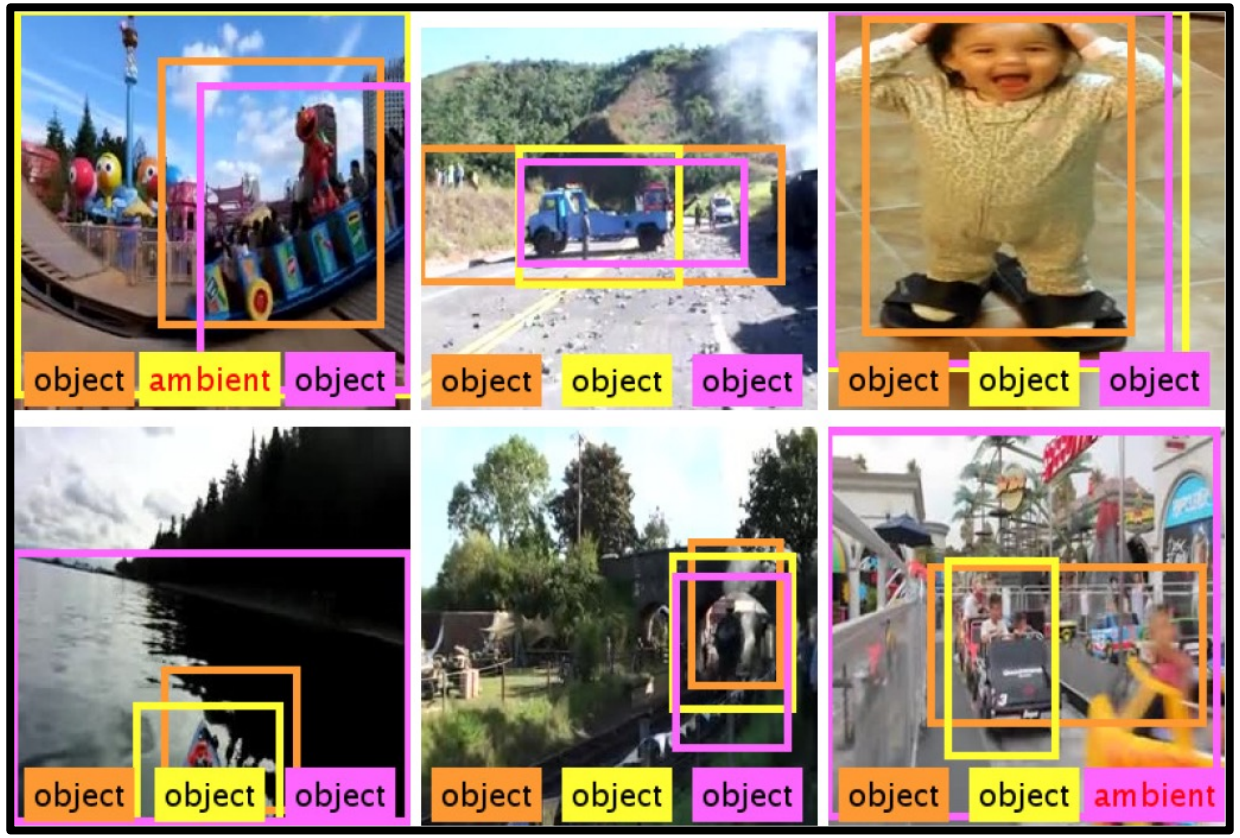
Unified architecture

- Semi-supervised learning case



Sound source localization dataset

(5k sounds x annotations from 3 subjects)

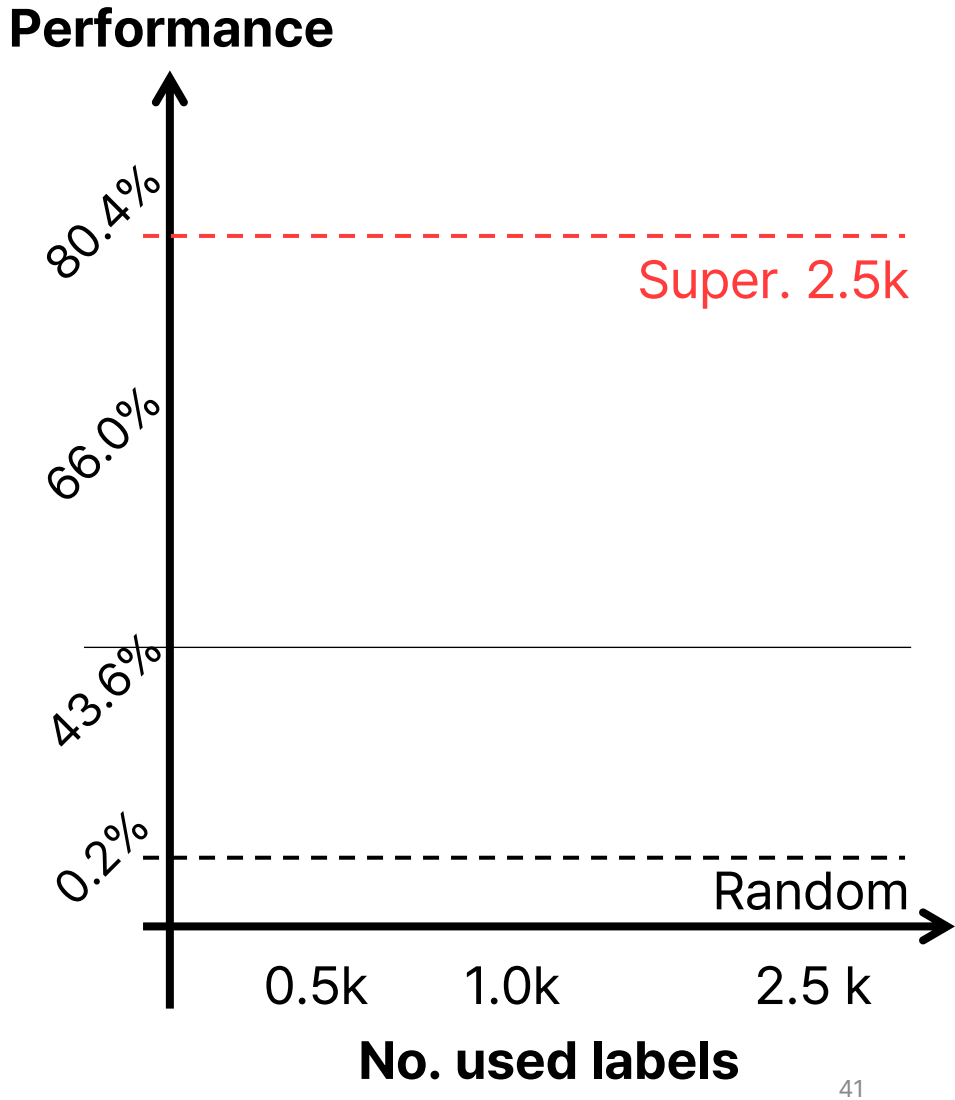


https://github.com/ardasnck/learning_to_localize_sound

Evaluation

- Supervised results

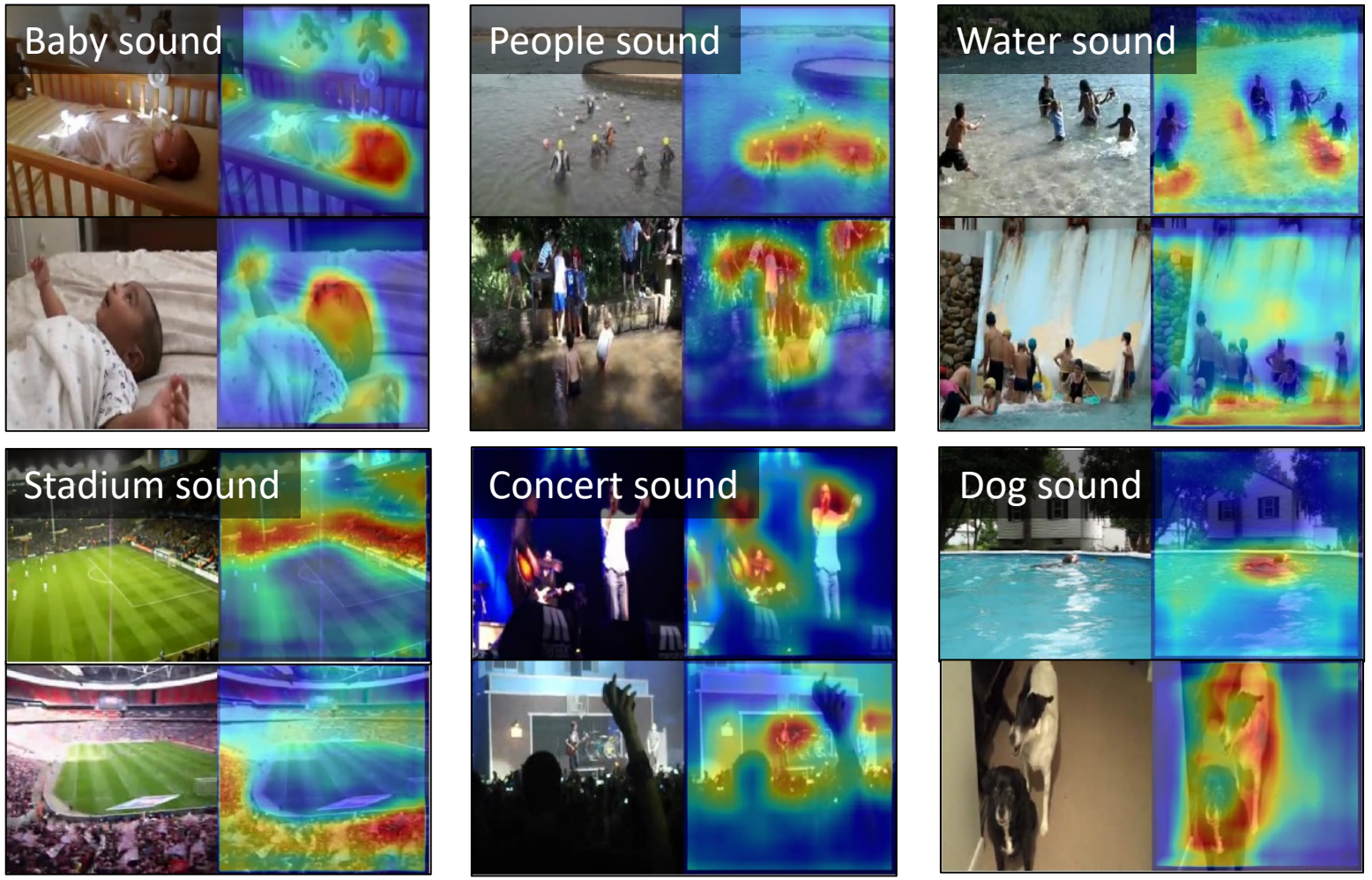


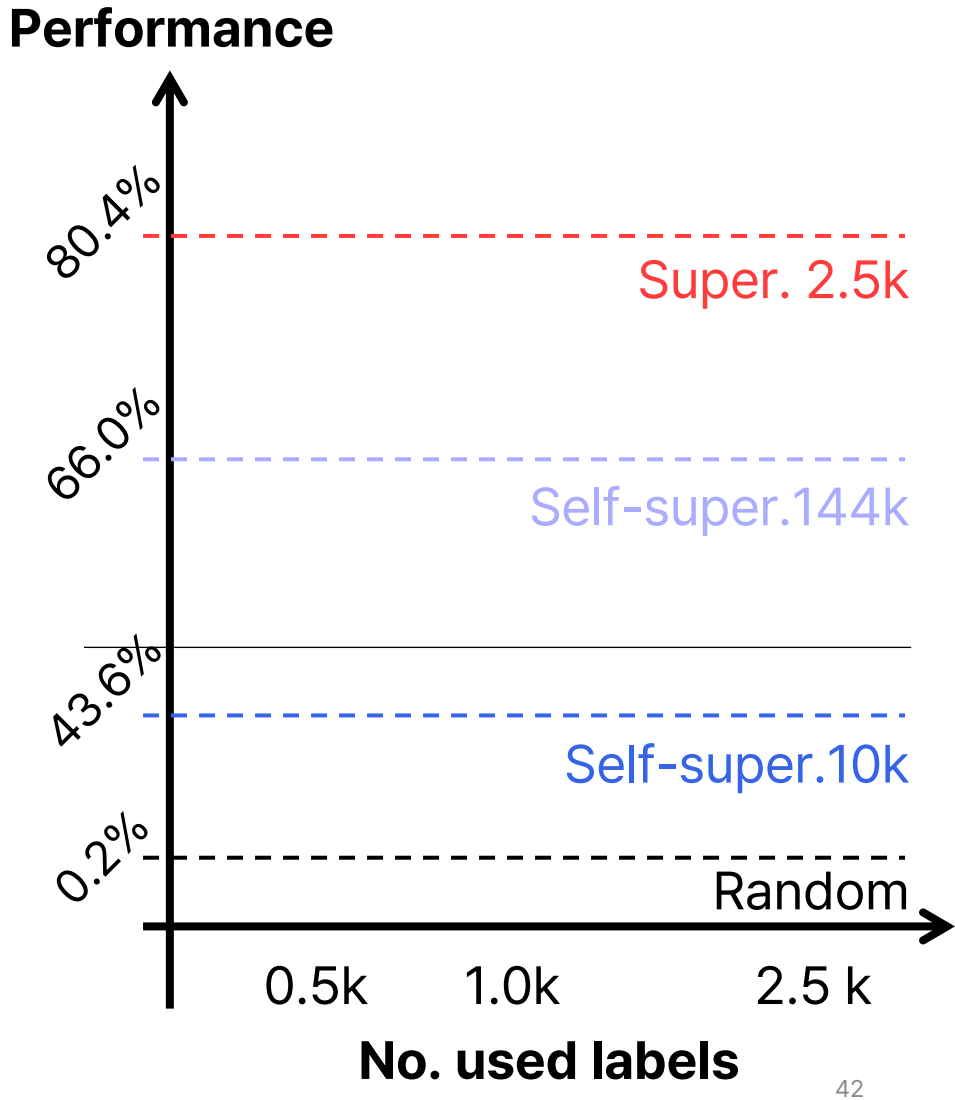


- Each frame is independently processed
- No motion or temporal cues are used
- No category information is used

Evaluation

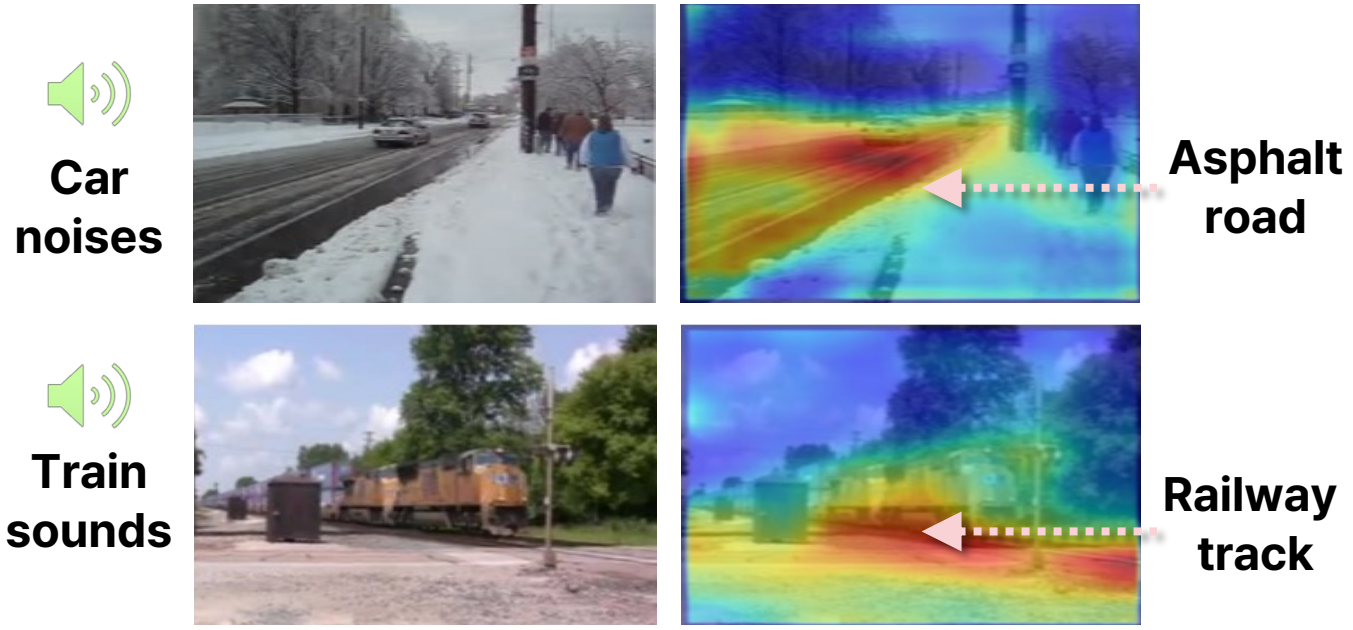
- Self-supervised results





A major failure type

- Failure case



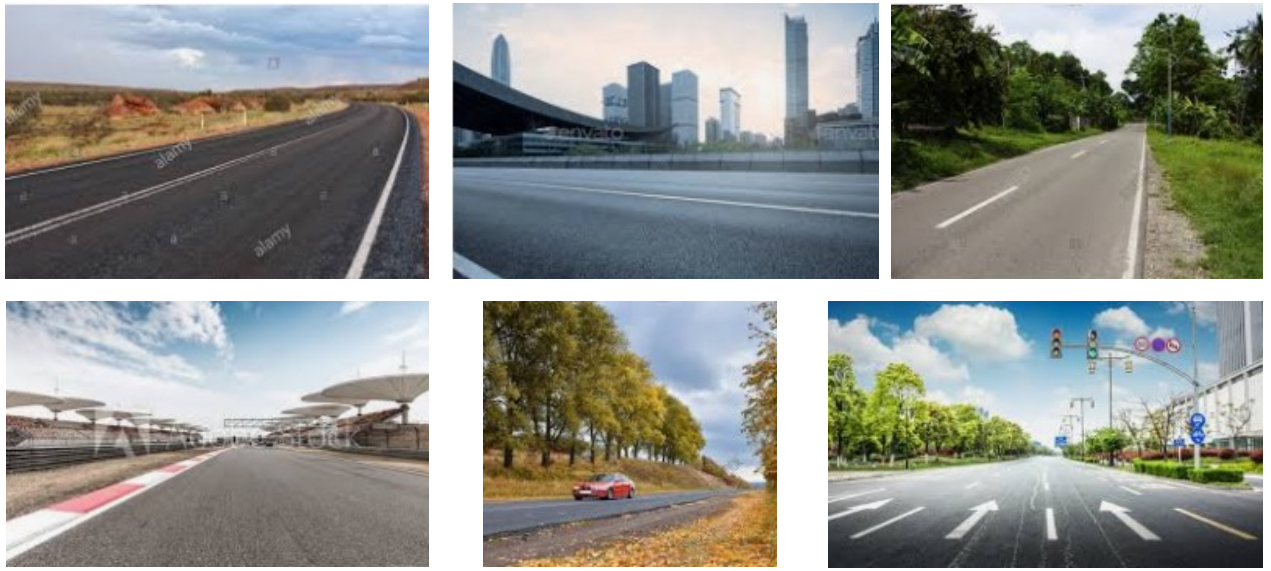
The model mistakenly learns that the sound are coming from the road

Failure cases induced from co-occurence

1. Visual appearance diversity (Car > Road) ⇒ Expose both at the same time
   ⇒ Neural Net takes the convenience
2. Correlation ≠ Causality



Car appearance



Road appearance

Senocak et al., Learning to Localize Sound Sources in Visual Scenes: Analysis and Applications, IEEE TPAMI 2021

Failure cases induced from co-occurence

- When Correlation≠Causality, humans & animals & ML models can believe wrong correlation as true

Without exact prior knowledge, learning cannot be corrected
[Ben-David]



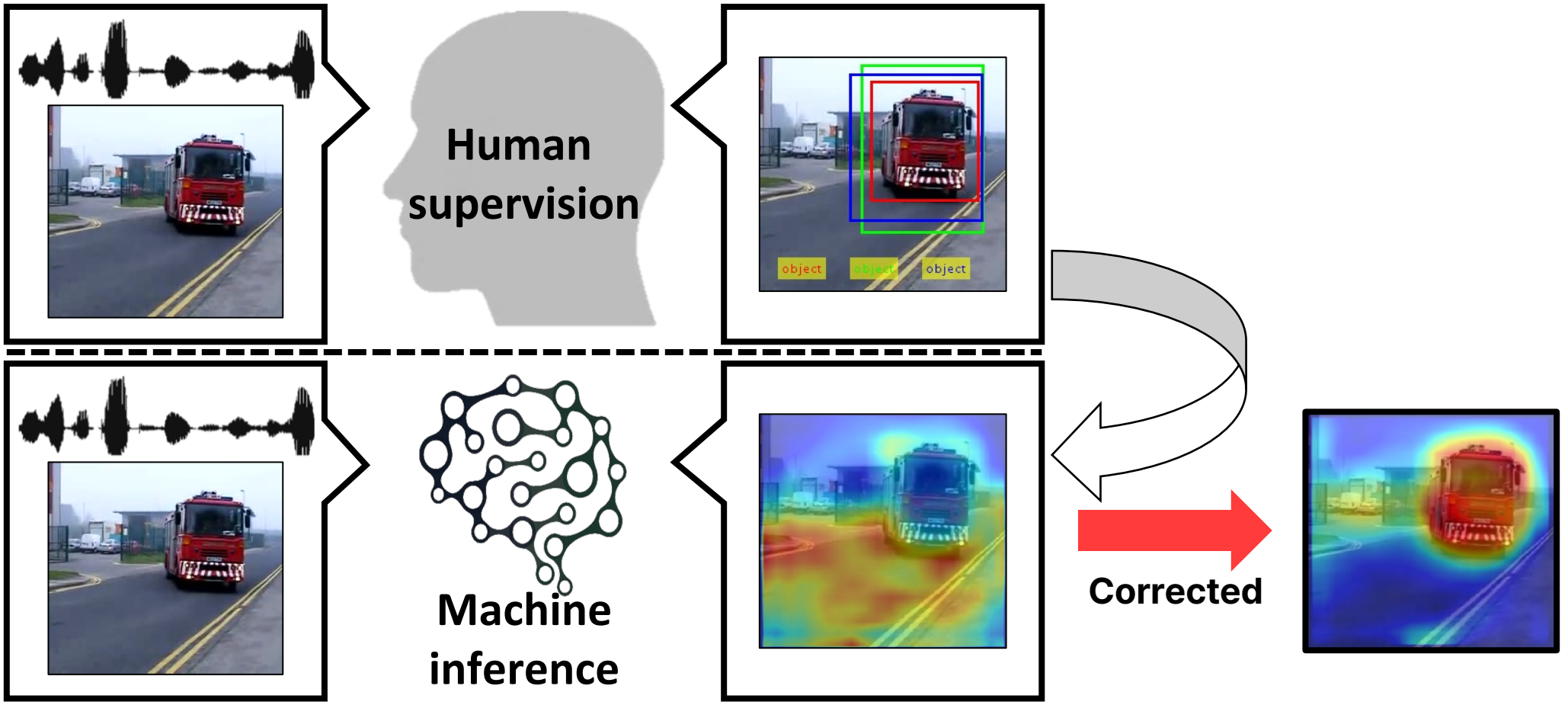


Senocak et al., Learning to Localize Sound Sources in Visual Scenes: Analysis and Applications, IEEE TPAMI 2021
Pigeon superstition phenomenon by Skinner
Shai Ben-David, Understanding Machine Learning: From Theory to Algorithms, 2003

Leverage a small amount of labels

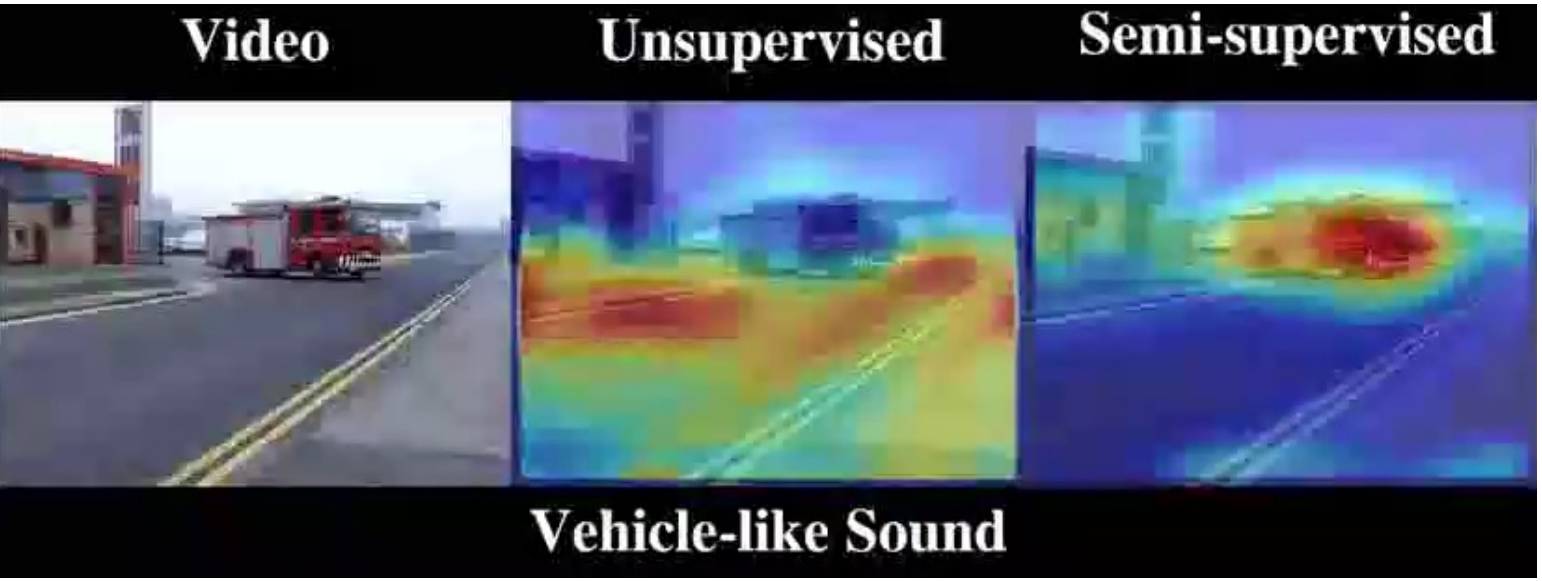
- Semi-supervised learning improves outcomes

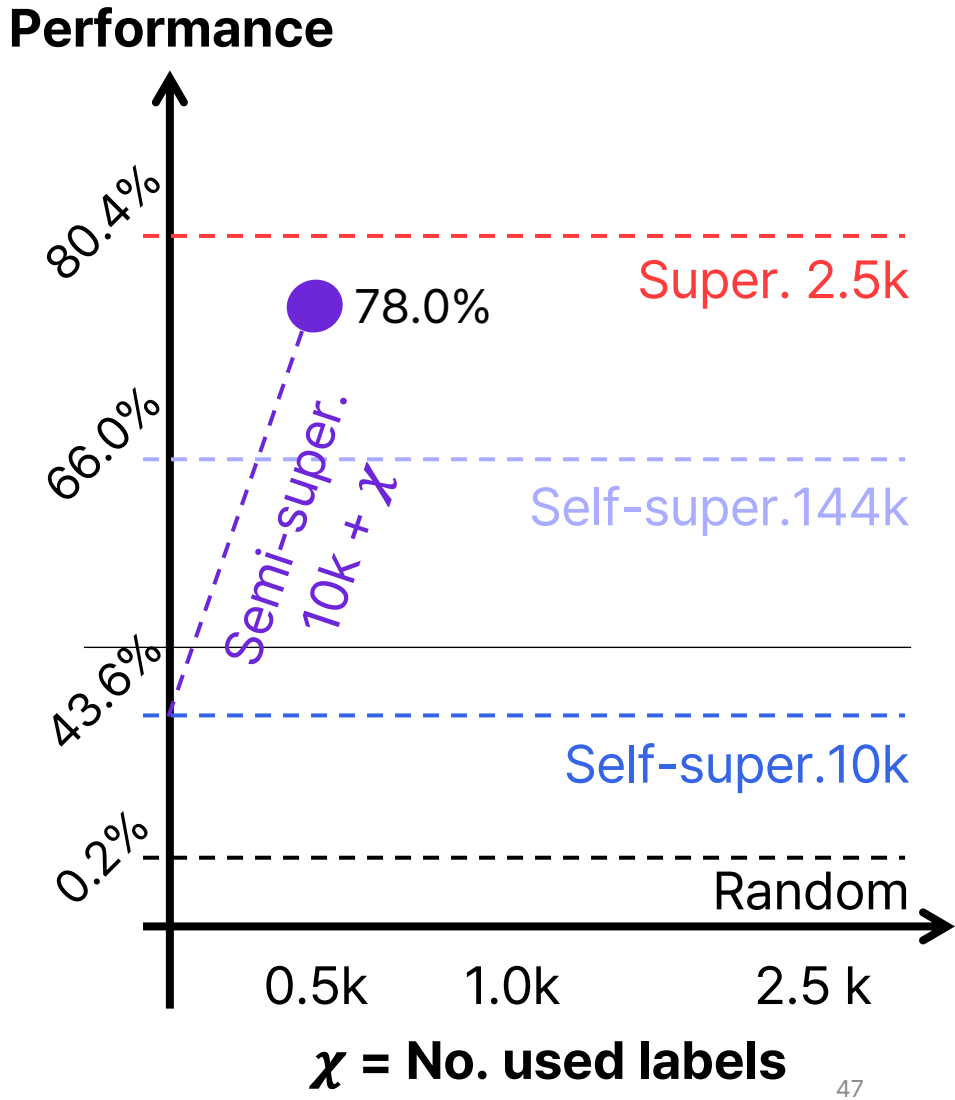
  by a small amount of human guidance



Evaluation

- Semi-supervised results



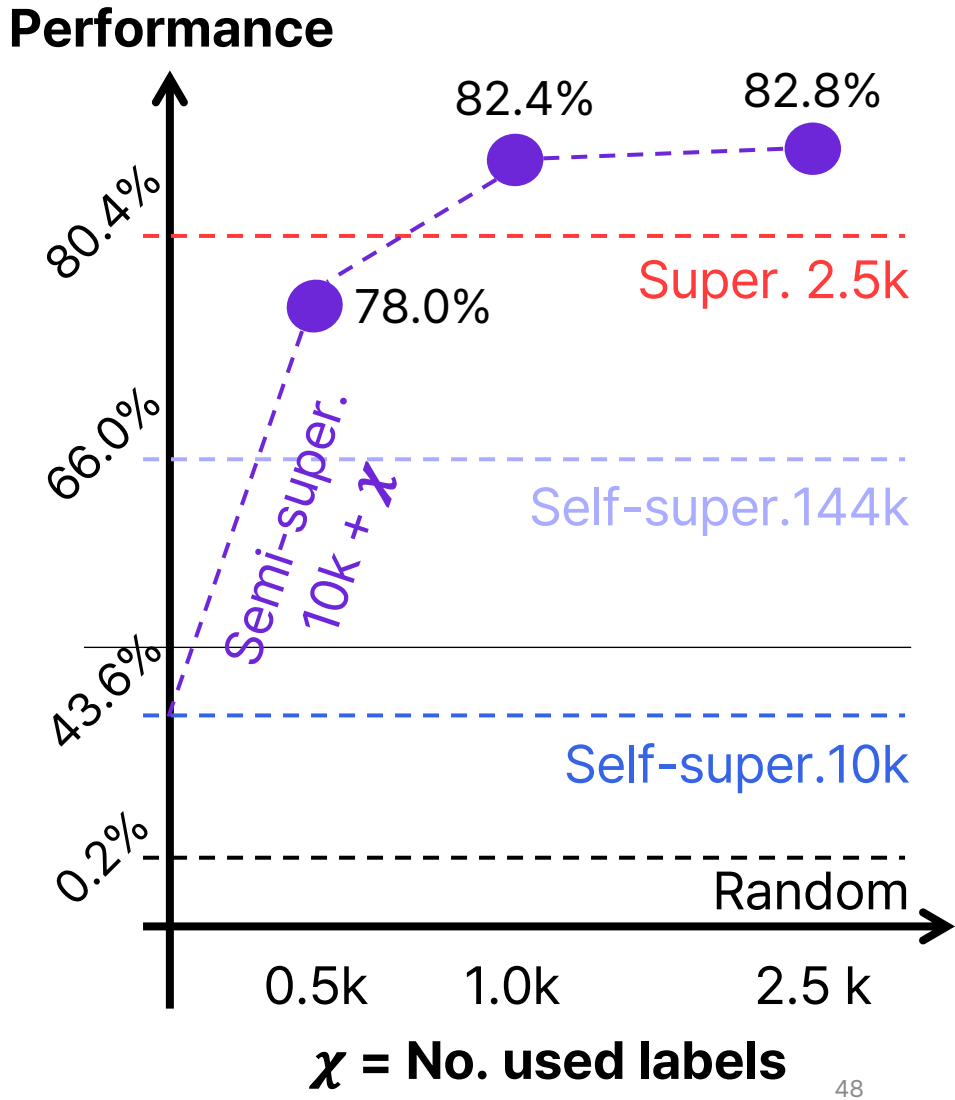


- Each frame is independently processed
- No motion or temporal cues are used
- No category information is used

Evaluation

- Semi-supervised results





- Each frame is independently processed
- No motion or temporal cues are used
- No category information is used

Labeled Data as Prior Information

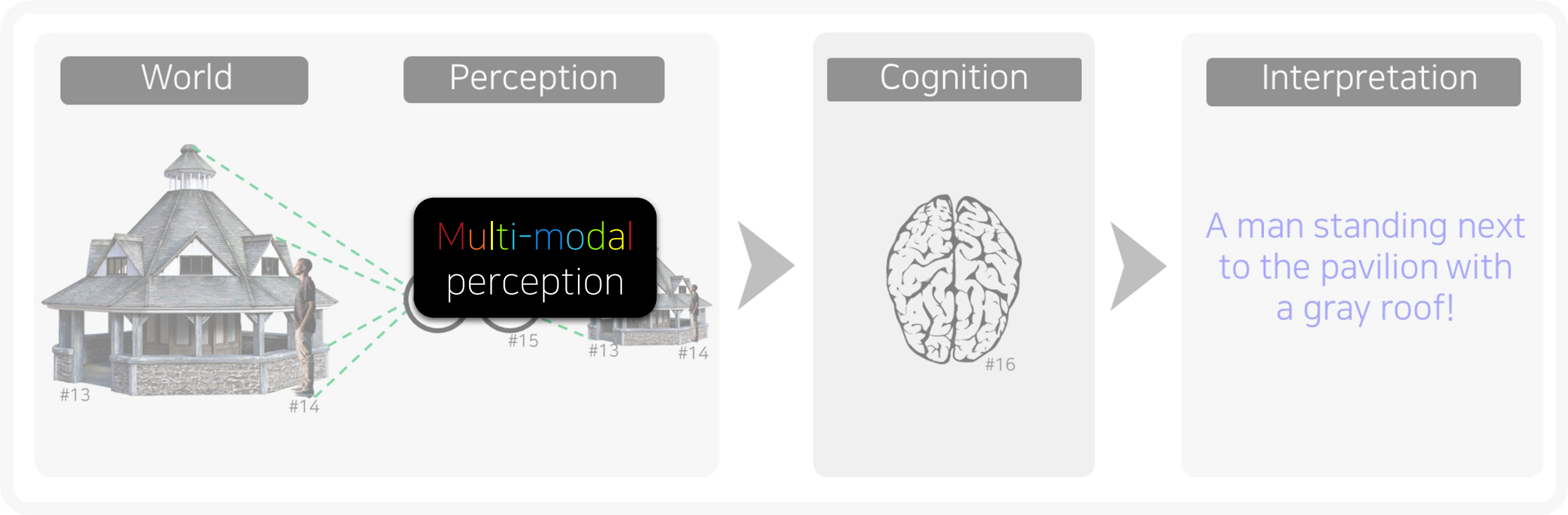
Multimodal perception system

cognition

perception

higher cognitive functions

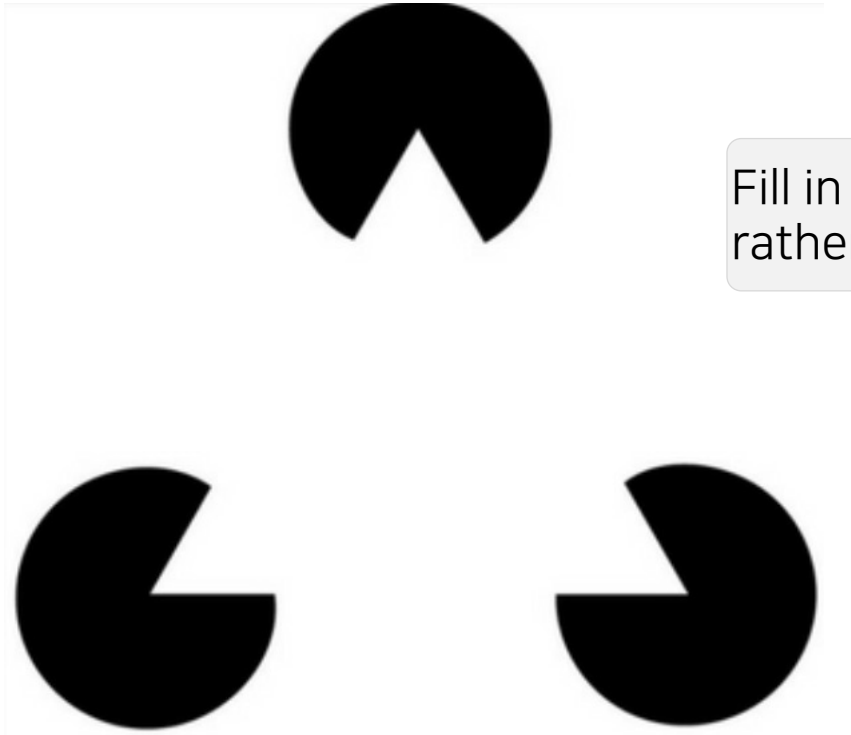
Other signals help compensate for missing or weak modalities



Human's scene understanding process

Human vision system

Gestalt Principle Humans perceive objects as wholes, rather than just a collection of parts

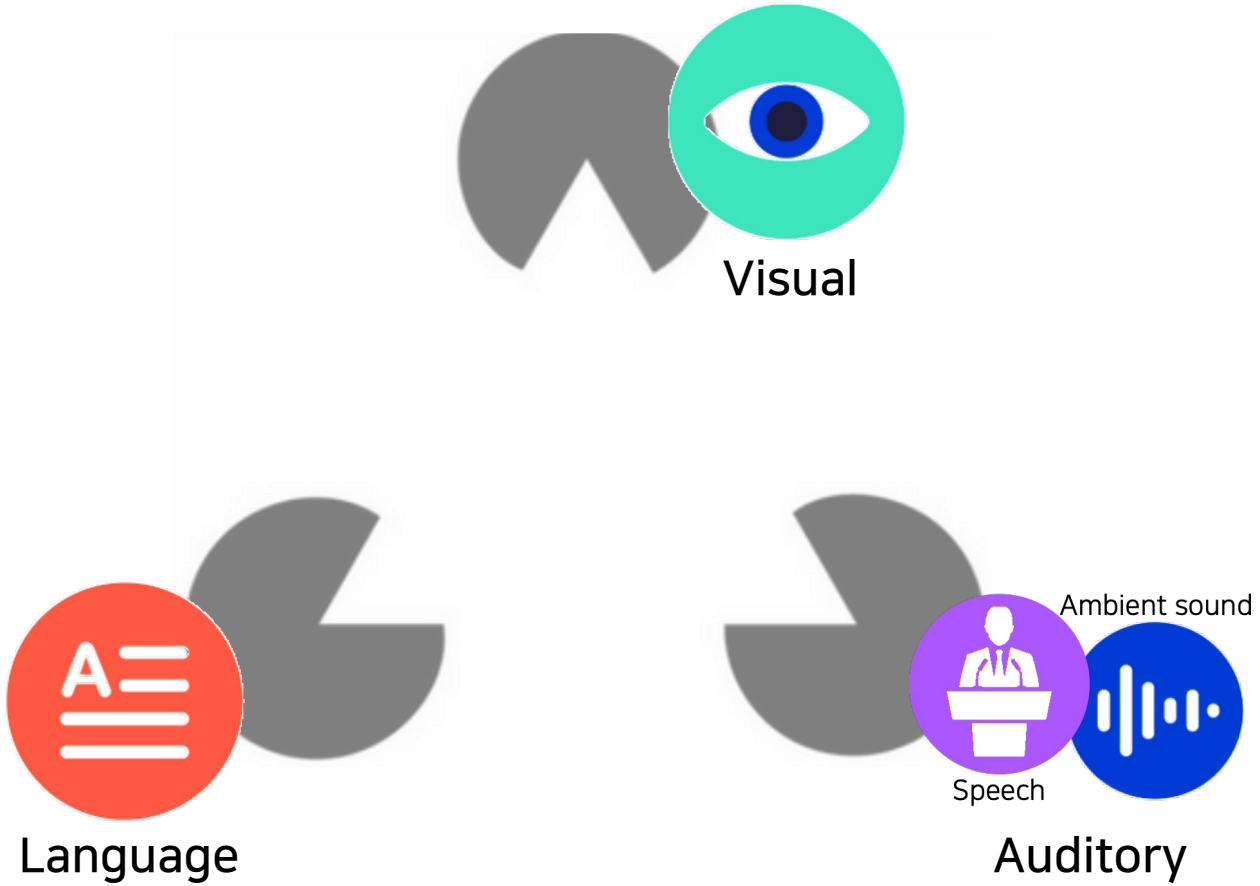


Fill in the gap by recognizing whole, rather than parts

An example of Gestalt's principle

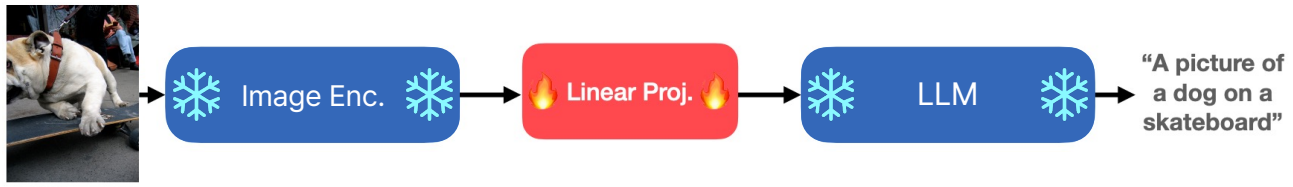
Multimodal perception system

Multi-modal Gestalt Fill in the gap by recognizing multi-modality to reveal the invisible



Recent Multi-modal Models

Common Design Pattern

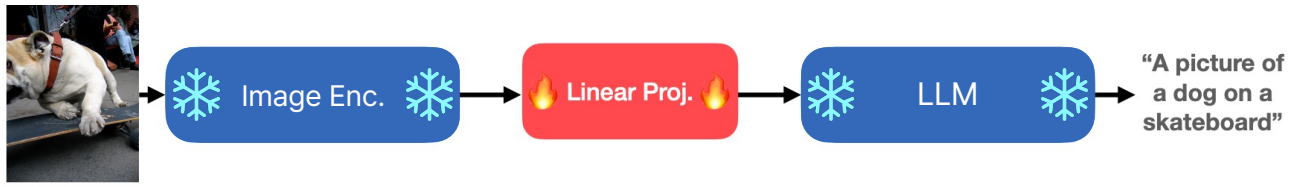


- Connect the pre-trained modules
  - Learning simple mapping module
  - Input is directly transferred to next module

Merullo et al. "Linearly mapping from image to text space." ICLR 2023
Liu et al, "Visual Instruction Tuning", NeurIPS 2023

Recent Multi-modal Models

Common Design Pattern

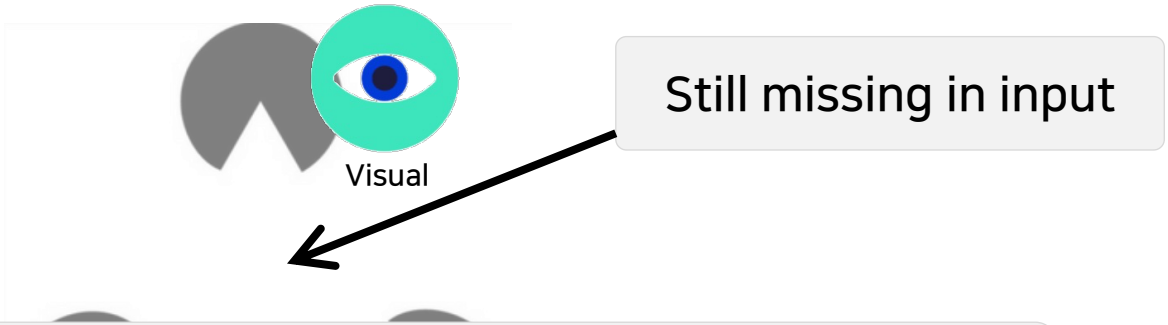


- Connect the pre-trained modules
  - Learning simple mapping module
  - Input is directly transferred to next module

Merullo et al. "Linearly mapping from image to text space." ICLR 2023
Liu et al, "Visual Instruction Tuning", NeurIPS 2023

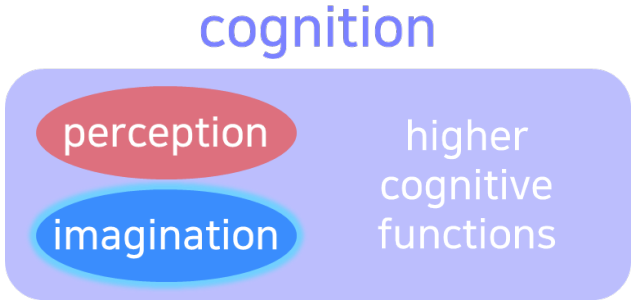
When information is missing in input

- ML models are “Garbage in, garbage out (GIGO)” programs
- ML models cannot learn without correlation btw. input & output
- Even another modality is given to a multimodal system,
  that signal might be weak (low SNR)

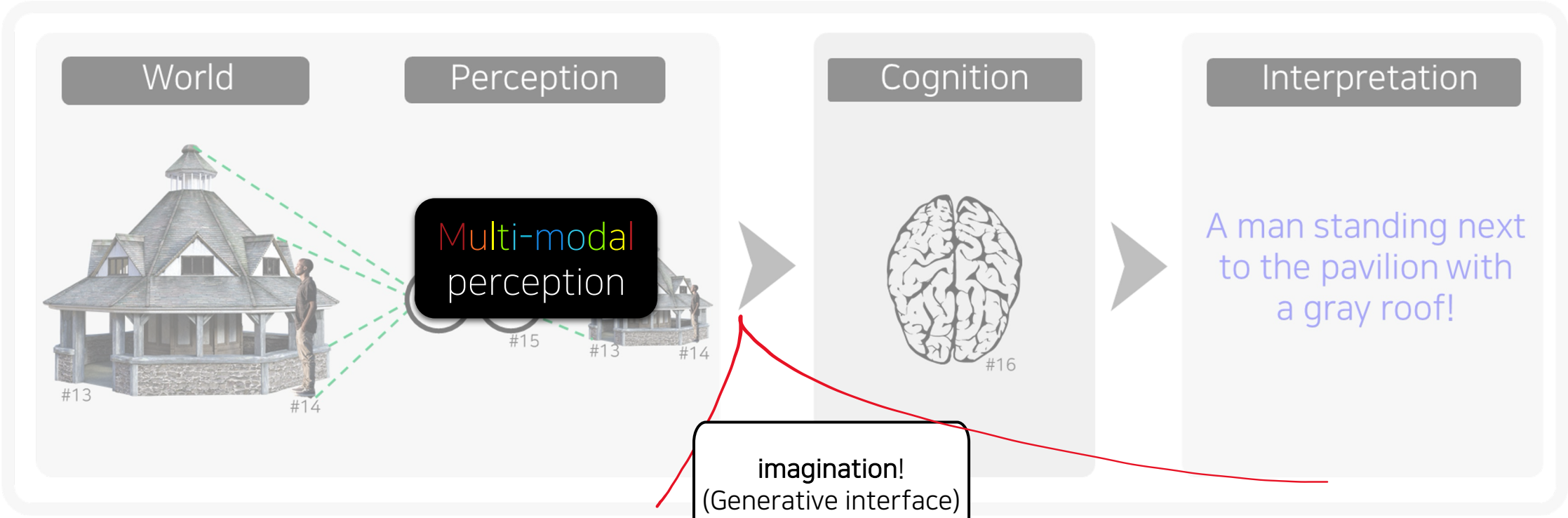


In most cases, your NNs are trained to capture ONLY apparent signals

Generative Machine Perception



Human intelligence = perception ∪ higher cognitive processes



The way to fill in missing in computer vision

- To overcome this, we need
  1. An *additional signal*

The way to fill in missing in computer vision

- To overcome this, we need
  1. An *additional signal*
  2. *Imagination* to complete missing information

The way to fill in missing in computer vision

- To overcome this, we need
  1. An *additional signal*
  2. *Imagination* to complete missing information
  3. The way to associate them

The way to fill in missing in computer vision

- To overcome this, we need
  1. An *additional signal* = **Sound**
  2. *Imagination* to complete missing information = **Generative model**
  3. The way to associate them = **Interface**

Generative interface

= Generative cross-modal translation to viz.

- Incomplete perception to visual perception
- Get a hint to develop an ultimate generative interface for machine perception

Pre-trained Model as Prior Information

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Also, extended to IJCV 2026

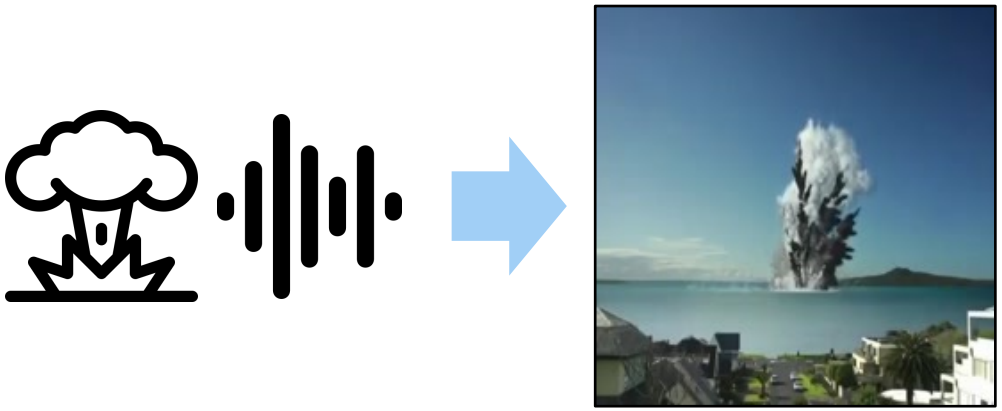
Sound to Visual Scene Generation by Audio-to-Visual Latent Alignment

Kim Sung-Bin[1] Arda Senocak[2] Hyunwoo Ha[1] Andrew Owens[3] Tae-Hyun Oh[1]

[1]POSTECH [2]KAIST [3]Univ. of Michigan

YONHAP NEWS AGENCY

YTN



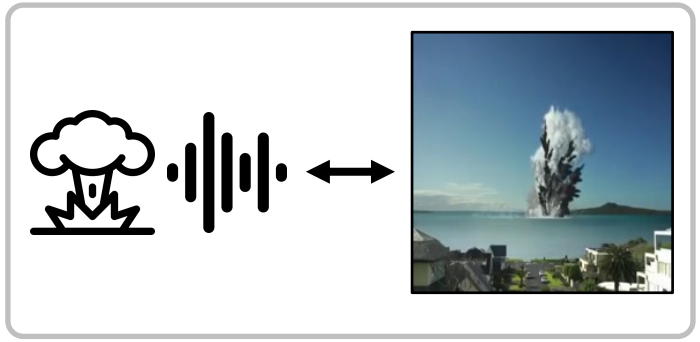
Audio event Visual event

Sound to image generation

Goal: learning to visualize the visual semantics embedded in sound

Sound to image generation

Goal: learning to visualize the visual semantics embedded in sound just by showing a lot of Youtube video



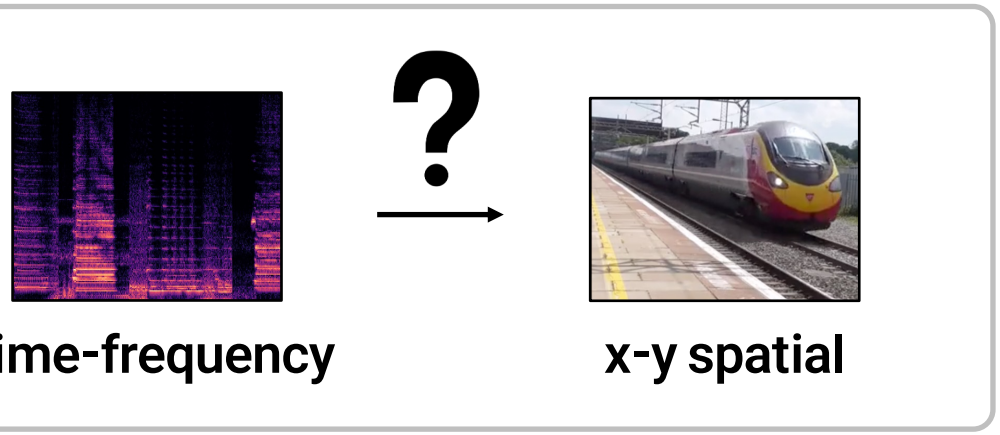
Audio-image pairs



Video

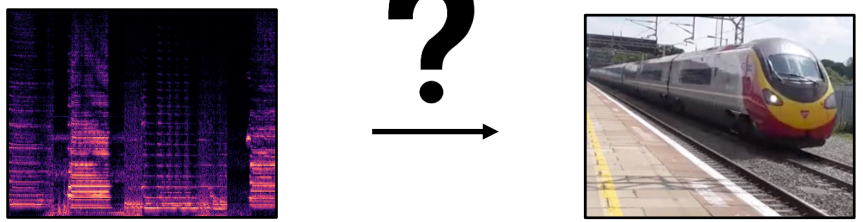
(co-occurring)

Challenges

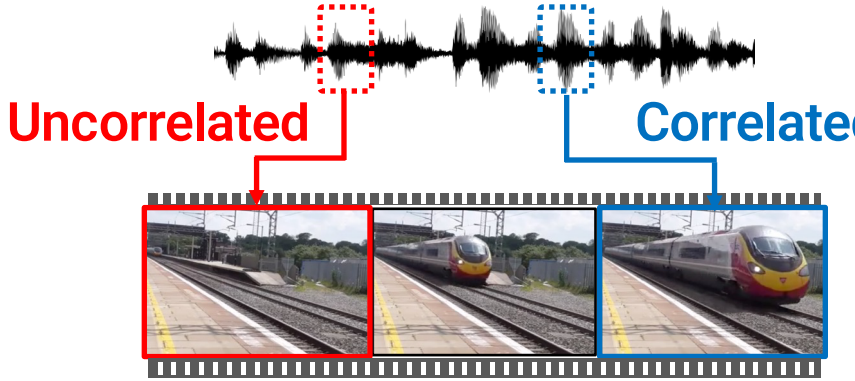


Audio-visual modality gap

Challenges

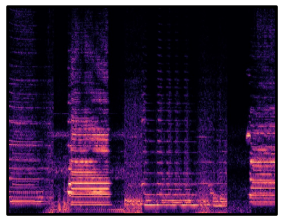


Audio-visual modality gap



Outliers in training data

Previous methods vs. Ours



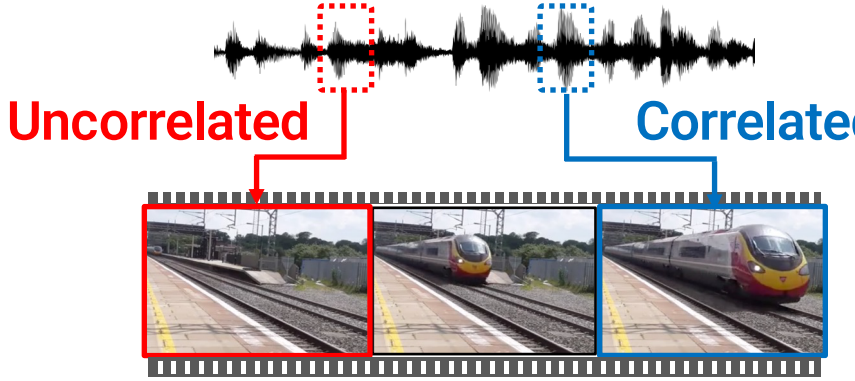
Time-frequency

?



x-y spatial

Audio-visual modality gap



Outliers in training data

Wan et al. Towards Audio to Scene Image Synthesis using Generative Adversarial Network, ICASSP, 2018
Fanzeres et al. Sound-to-Imagination, Arxiv, 2022

Previous methods vs. Ours

Audio-visual modality gap

Outliers in the dataset

Wan et al. Towards Audio to Scene Image Synthesis using Generative Adversarial Network, ICASSP, 2018
Fanzeres et al. Sound-to-Imagination, Arxiv, 2022

Previous methods vs. Ours

Audio-visual modality gap

Previous methods

Direct end2end learning

vs.

Ours

Module-based curriculum learning

Outliers in the dataset

Wan et al. Towards Audio to Scene Image Synthesis using Generative Adversarial Network, ICASSP, 2018
Fanzeres et al. Sound-to-Imagination, Arxiv, 2022

Design principle

- Mental imagery in blind people

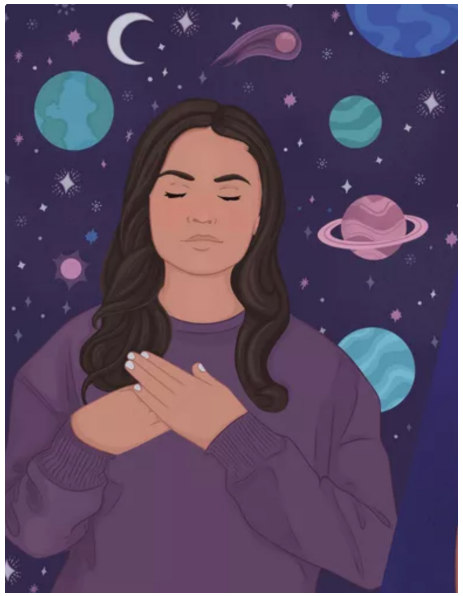
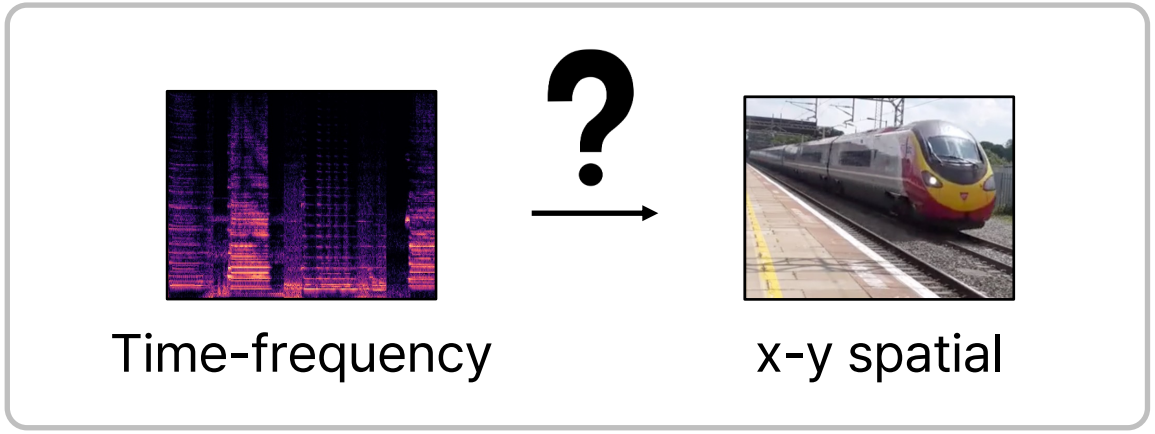


Image source: Verywell / Zoe Hansen

Becoming-blind still have visual mental images but *blind-born* don't have.

Cattanneo et al., "Imagery and spatial processes in blindness and visual impairment," Neuroscience & Biobehavioral Reviews, 2008
H Burton, Visual Cortex Activity in Early and Late Blind People, Journal of Neuroscience, 2003

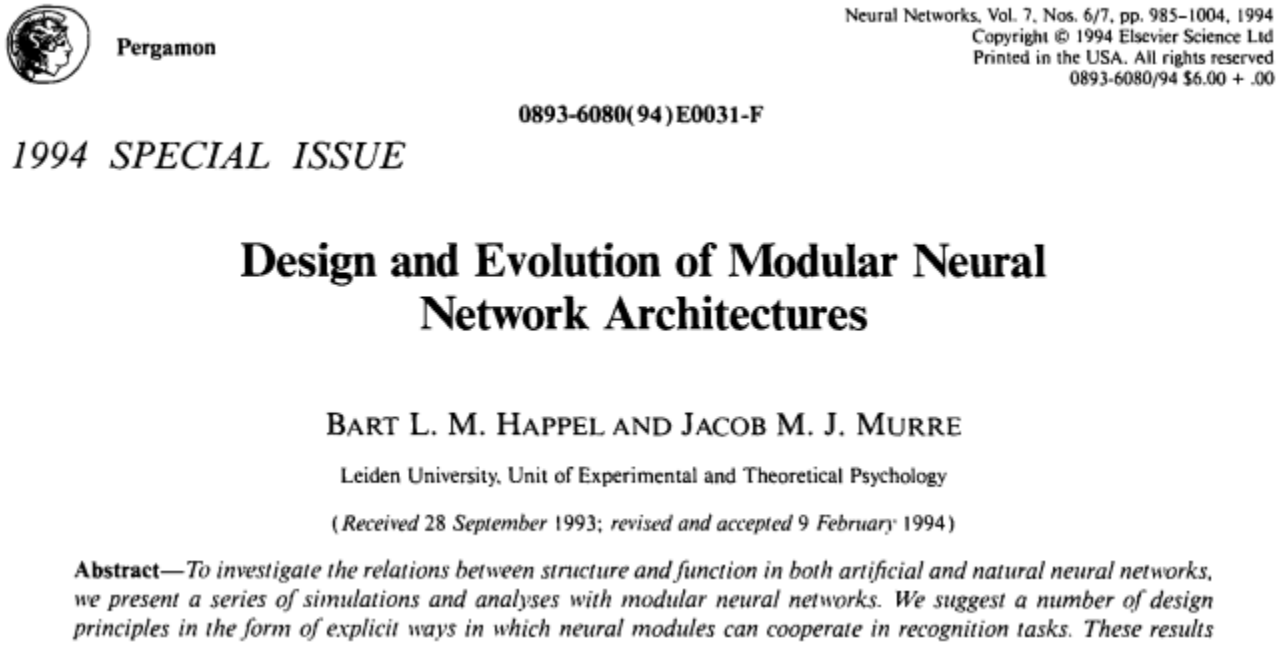
Design principle



Audio-visual modality gap

Establish clean visual knowledge first,
and then capture visual related information from audio

Design principle



Pergamon

Neural Networks, Vol. 7, Nos. 6/7, pp. 985–1004, 1994
Copyright © 1994 Elsevier Science Ltd
Printed in the USA. All rights reserved
0893-6080/94 $6.00 + .00

0893-6080(94)E0031-F

1994 SPECIAL ISSUE

Design and Evolution of Modular Neural Network Architectures

BART L. M. HAPPEL AND JACOB M. J. MURRE

Leiden University, Unit of Experimental and Theoretical Psychology

(*Received* 28 *September* 1993; *revised and accepted* 9 *February* 1994)

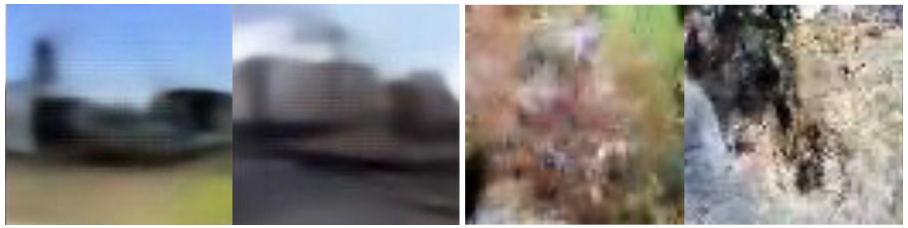
Abstract—*To investigate the relations between structure and function in both artificial and natural neural networks, we present a series of simulations and analyses with modular neural networks. We suggest a number of design principles in the form of explicit ways in which neural modules can cooperate in recognition tasks. These results*

Design the whole system module-by-module, then integrate

Previous methods vs. Ours

Audio-visual modality gap

Previous methods



Train Dog

vs.

Ours



Train Dog

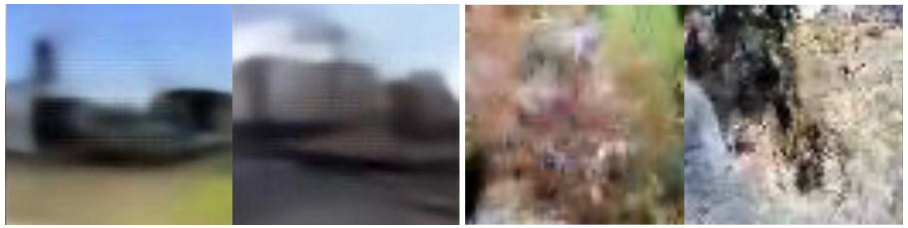
Outliers in the dataset

Fanzeres et al. Sound-to-Imagination, Arxiv, 2022
Wan et al. Towards Audio to Scene Image Synthesis using Generative Adversarial Network, ICASSP, 2018

Previous methods vs. Ours

Audio-visual modality gap

Previous methods



Train Dog

vs.

Ours



Train Dog

Outliers in the dataset

Previous methods

Random selection

vs.

Ours

Audio-visual correlation check

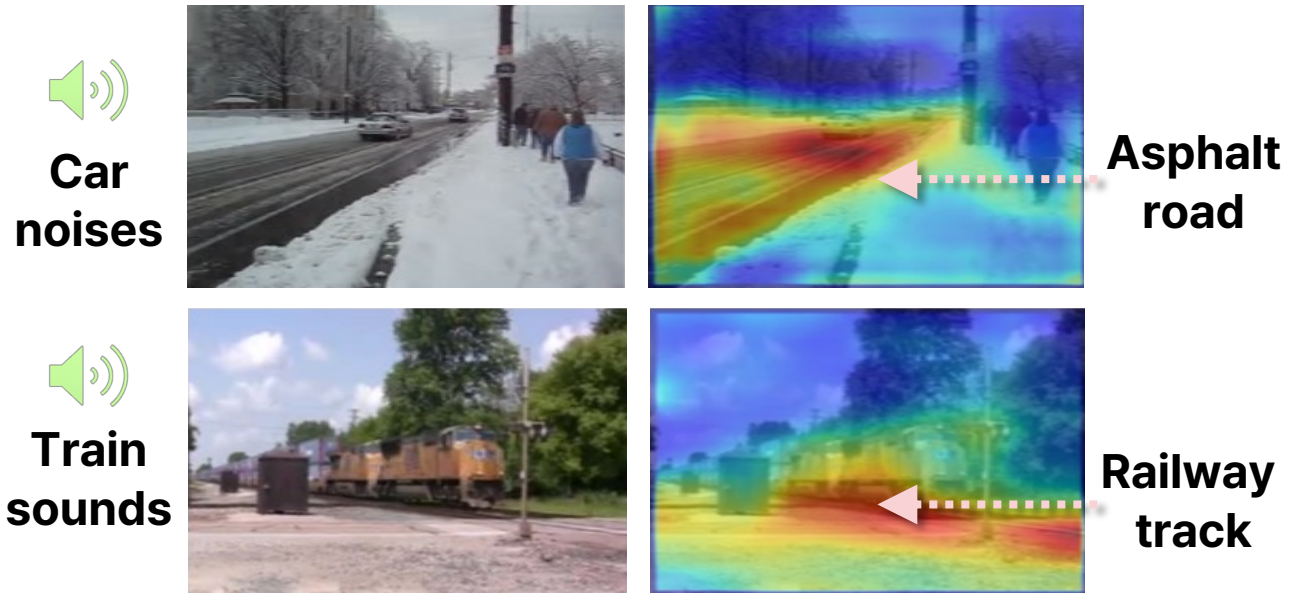
Wan et al. Towards Audio to Scene Image Synthesis using Generative Adversarial Network, ICASSP, 2018
Fanzeres et al. Sound-to-Imagination, Arxiv, 2022

Failure cases by outlier cases

Common failure cases in audio-visual self-supervised learning



Outliers in training data



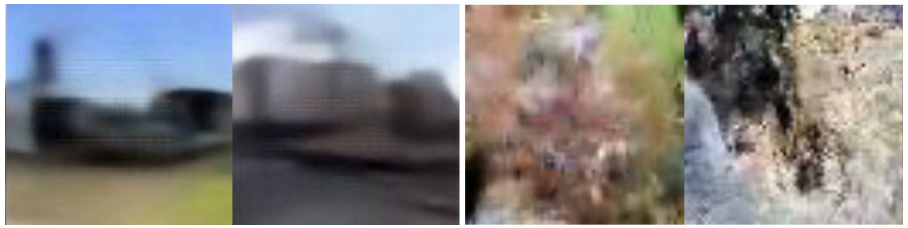
Semantically unmatched results in sound source localization

Senocak et al., Learning to Localize Sound Sources in Visual Scenes: Analysis and Applications, IEEE TPAMI 2021

Previous methods vs. Ours

Audio-visual modality gap

Previous methods



Train Dog

vs.

Ours



Train Dog

Outliers in the dataset

Previous methods

5-9 categories (restricted domain)

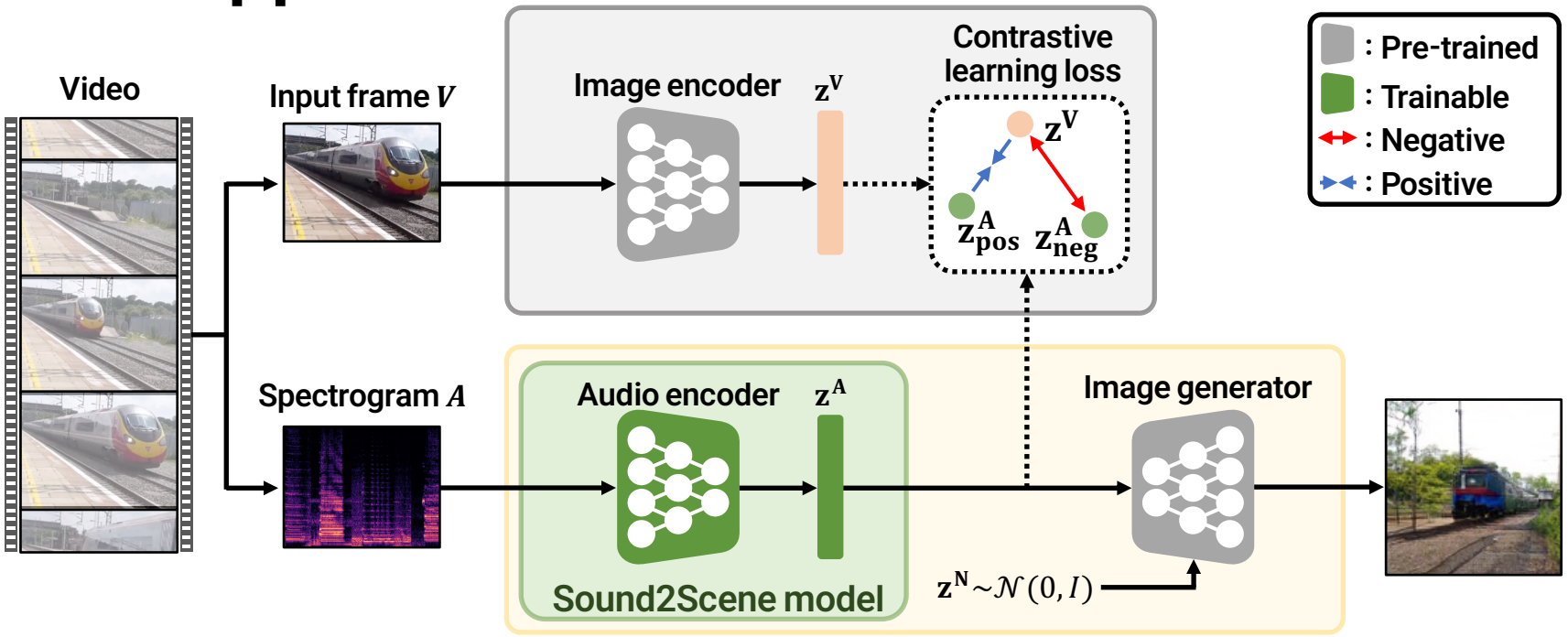
vs.

Ours

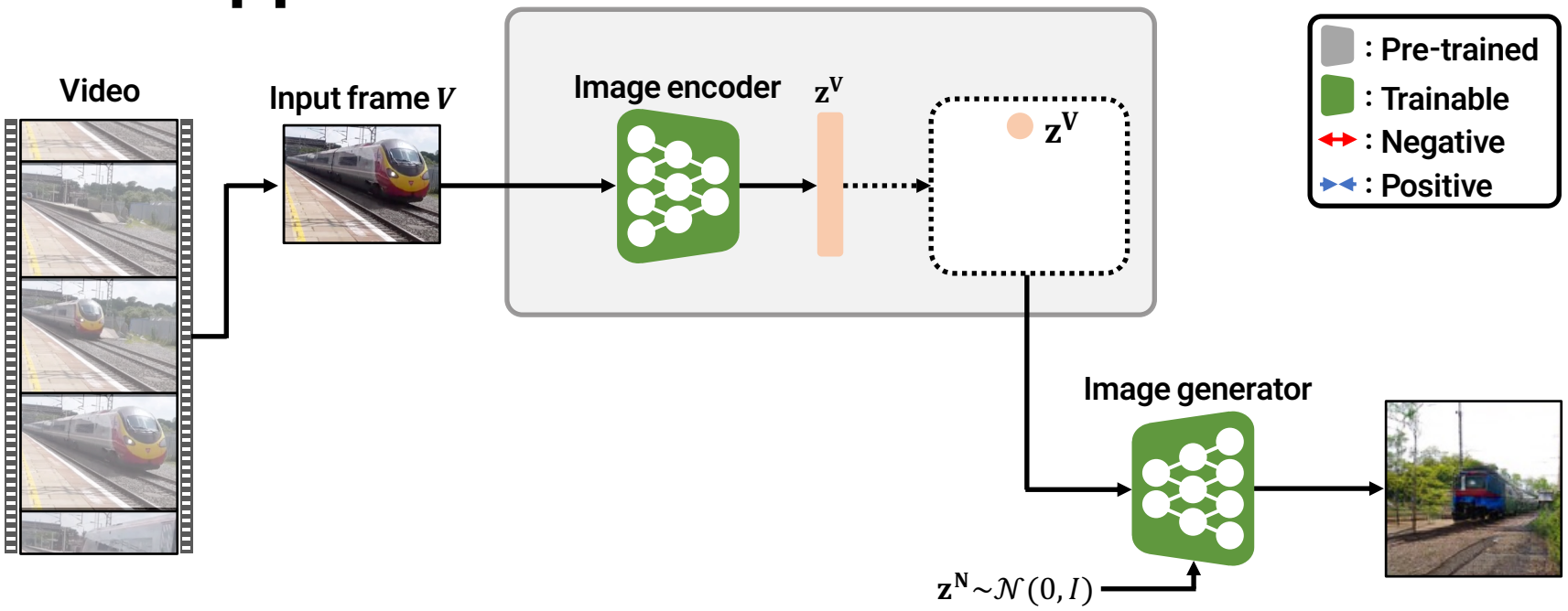
Any categories

Fanzeres et al. Sound-to-Imagination, Arxiv, 2022
Wan et al. Towards Audio to Scene Image Synthesis using Generative Adversarial Network, ICASSP, 2018

Module-based approach

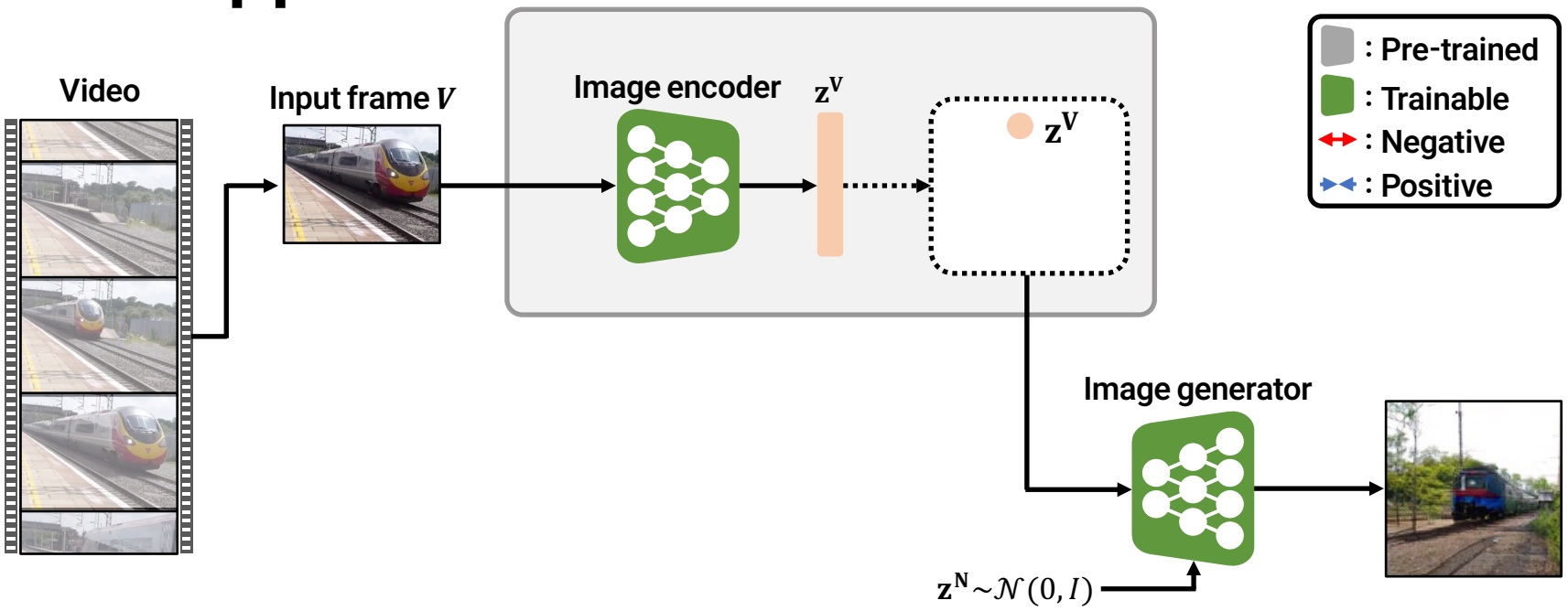


Module-based approach



Train a strong image encoder and generator separately

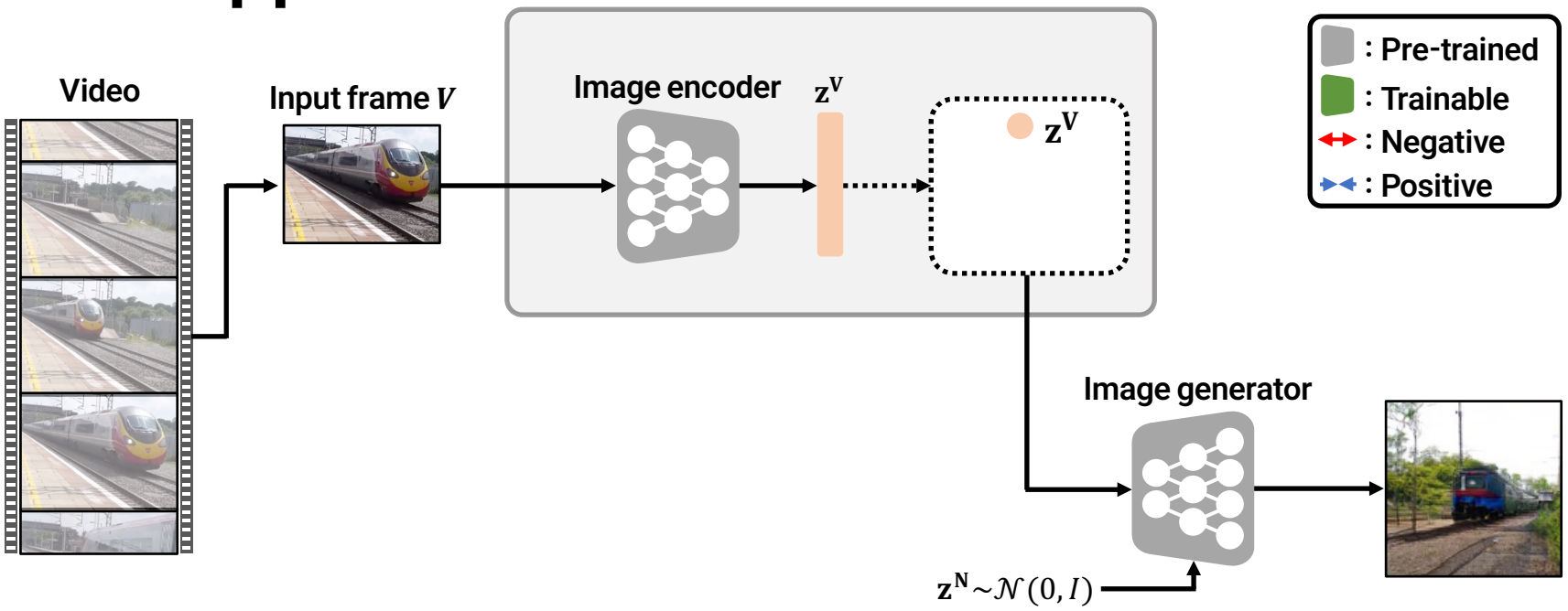
Module-based approach



Train a strong image encoder and generator separately

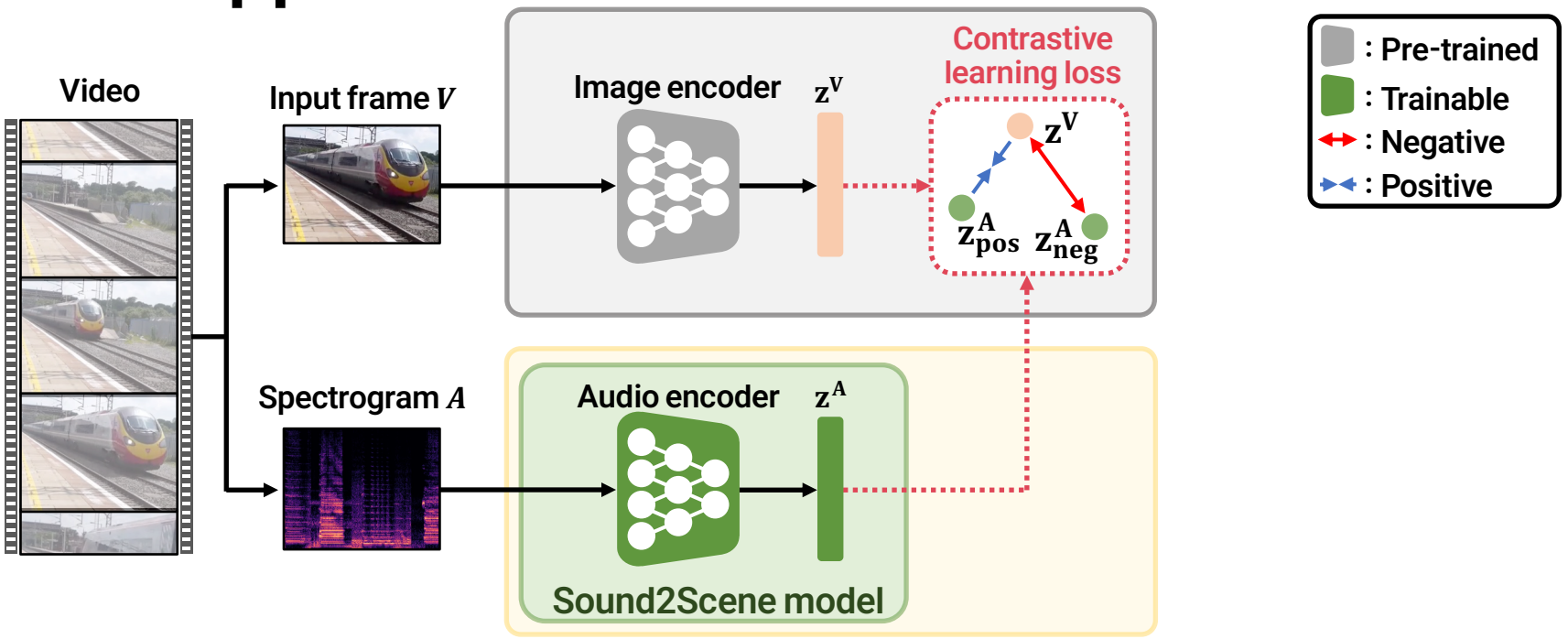
Build visual embedding space and expressive image generator

Module-based approach



Train a strong image encoder and generator separately (freeze)

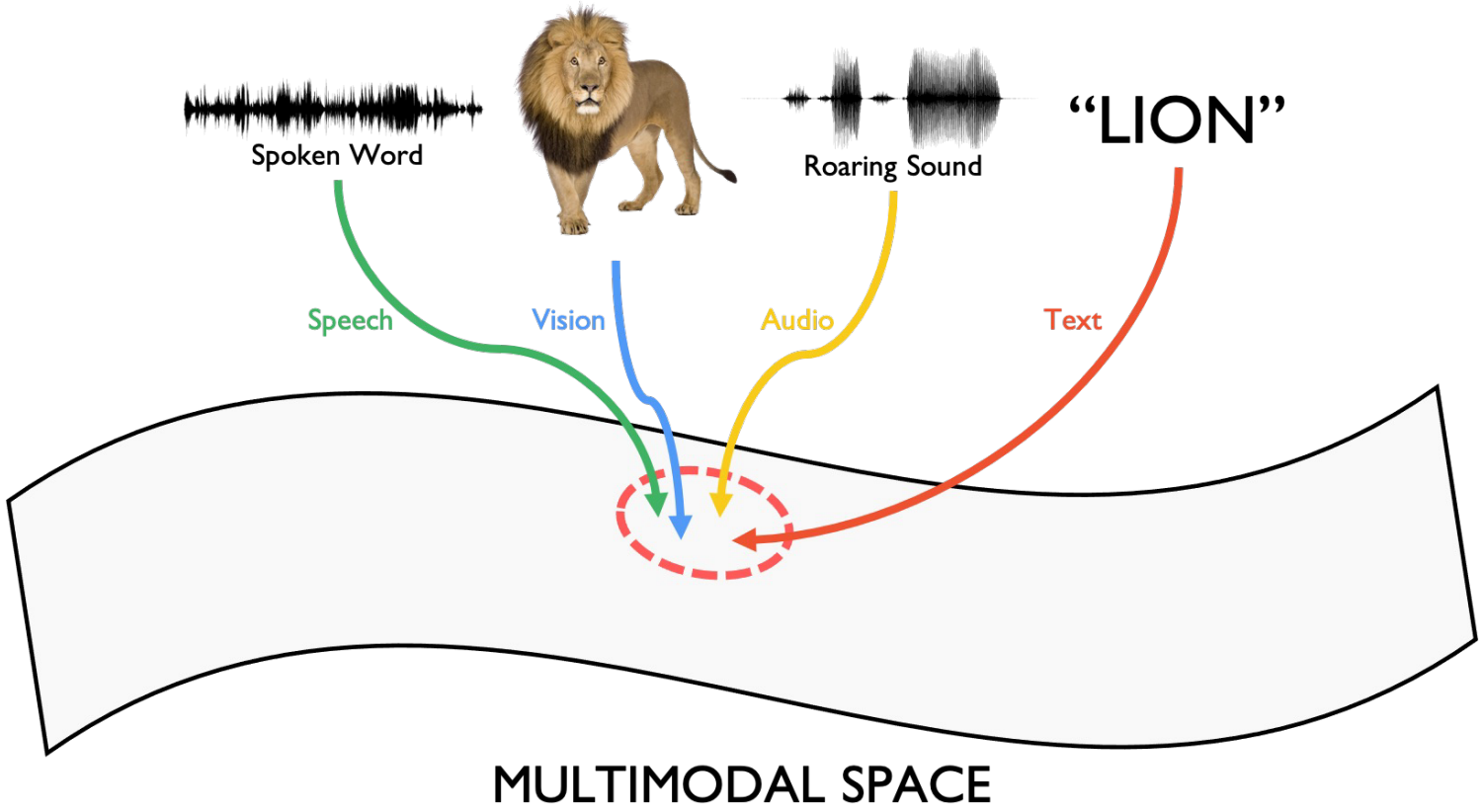
Module-based approach



Train a strong image encoder and generator separately (freeze)

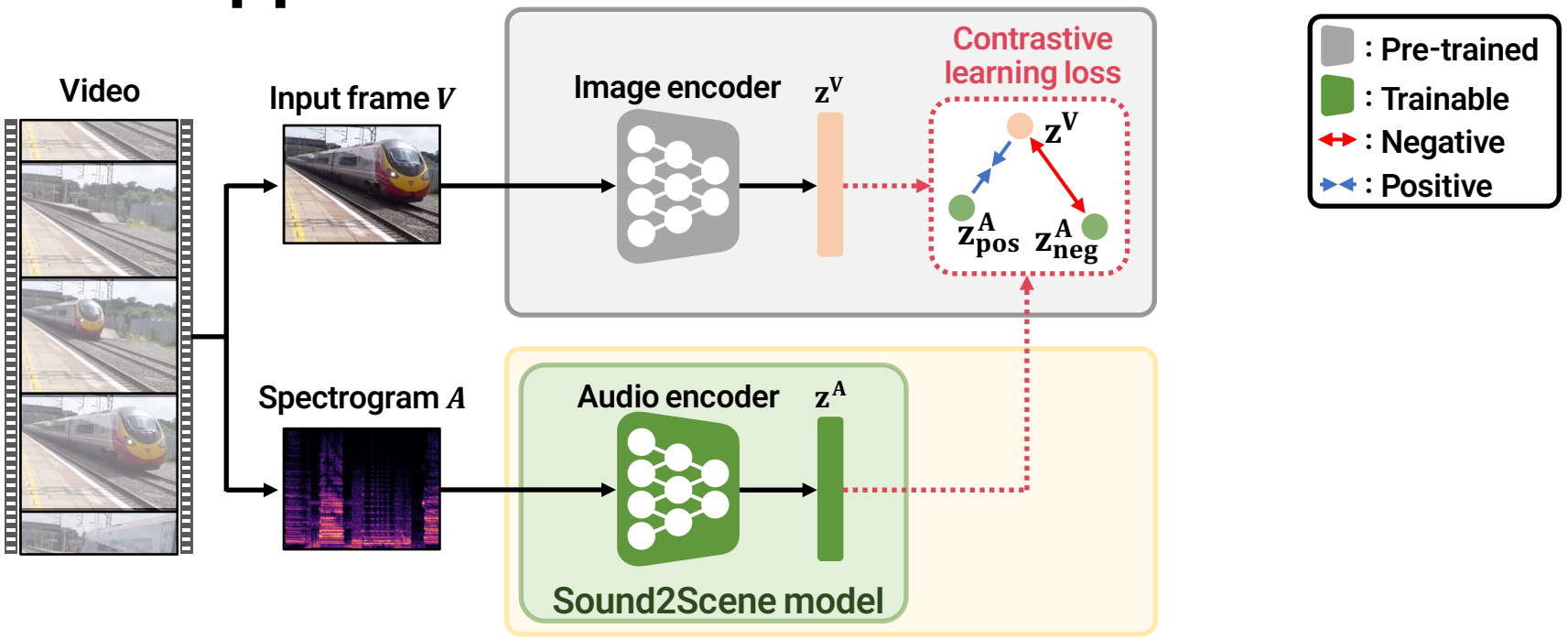
Train audio encoder to align audio-to-visual features

Multisensory Fusion for Concept Learning



GA Calvert. Crossmodal Processing in the Human Brain: Insights from Functional Neuroimaging Studies, Cerebral Cortex, 2001
Quiroga et. al. Invariant Visual Representation by Single Neurons in the Human Brain, Nature, 2005
Quiroga et. al. Explicit Encoding of Multimodal Percepts by Single Neurons in the Human Brain, Current Biology, 2009
Choksi et. al. Multimodal Neural Networks Better Explain Multivoxel Patterns in the Hippocampus, NeurIPS Workshop, 2021
Goh et. al. Multimodal Neurons in Artificial Neural Networks, OpenAI, 2021

Module-based approach

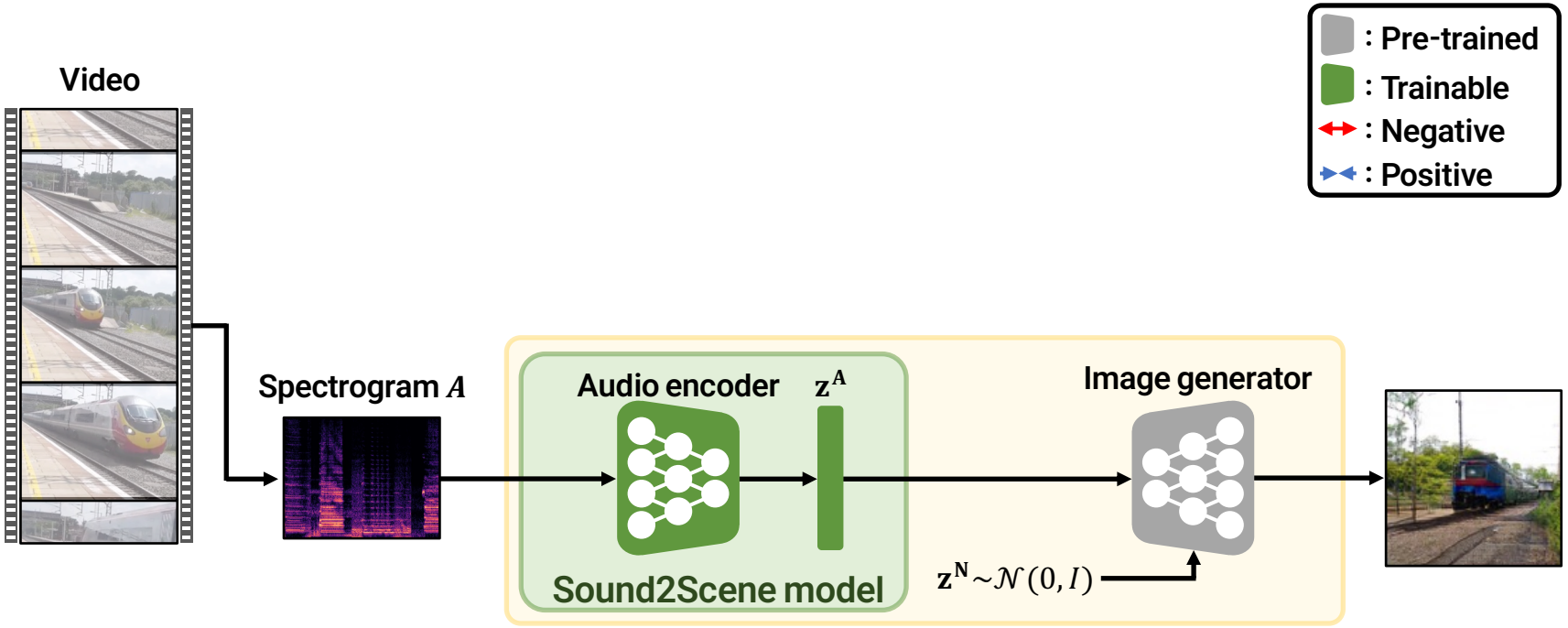


Train a strong image encoder and generator separately (freeze)

Train audio encoder to align audio-to-visual features

Enrich audio features with visual-centric information

Module-based approach



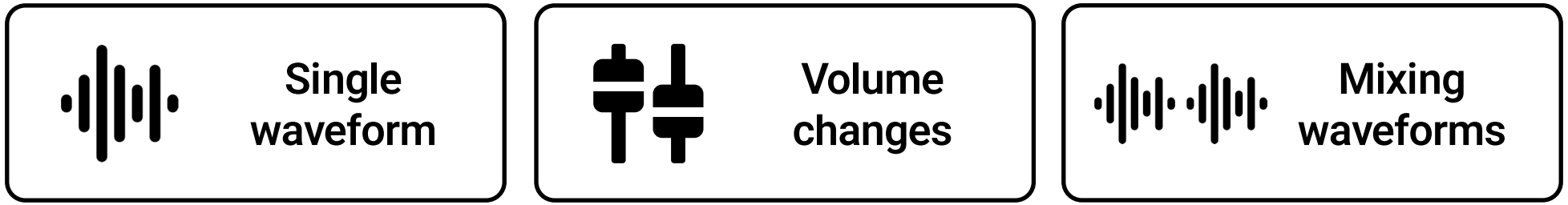
Train a strong image encoder and generator separately (freeze)

Train Sound2Scene to align audio-to-visual features

Feed audio features to frozen image generator

Qualitative result

Generate images from diverse sound characteristics



Result - Single waveform



Our model understands composition of sounds



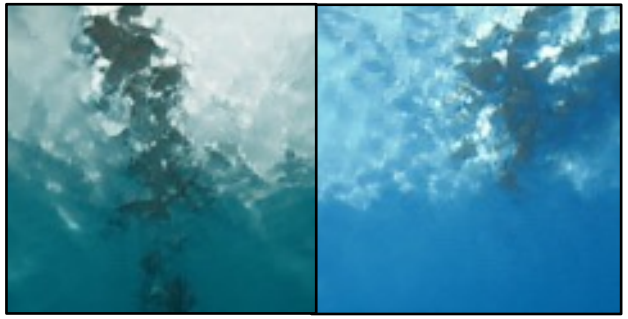
Dog barking + Water Flowing



Train Whistling + Skiing



Bird Singing + Skiing



Underwater + Explosion



Train Whistling + Hail

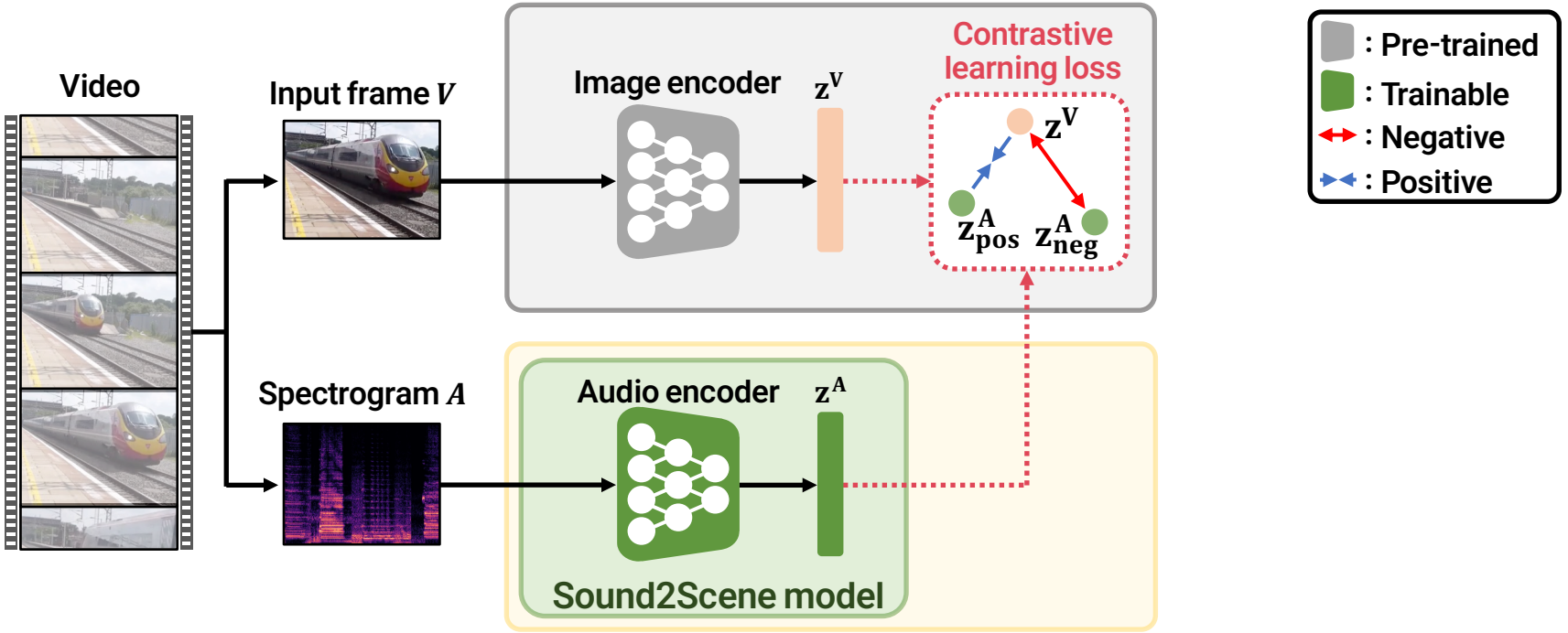


Bird Singing + Hail

Our model understands what volume change means



Applications



Can leverage audio-visual shared embedding space

Image-audio conditioned image generation

Given



Tractor

Generated image



Given



Explosion

Generated image

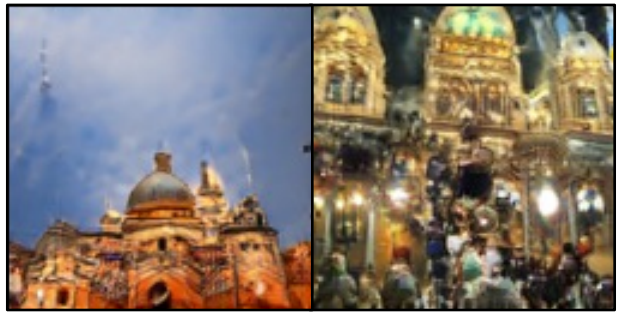


Given



Cheering

Generated image



Given



Skiing

Generated image

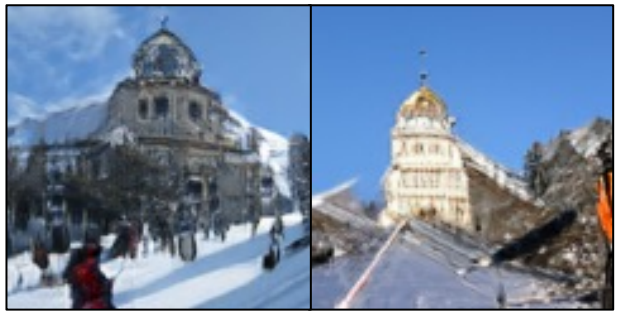
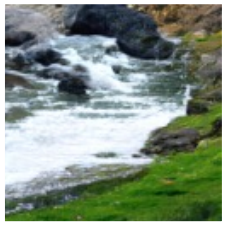


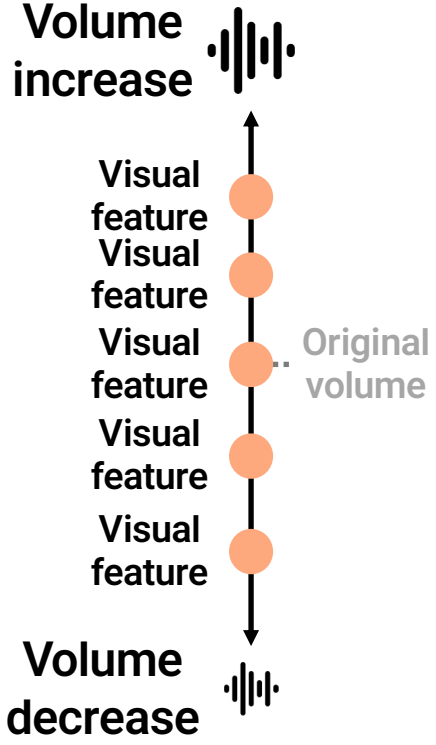
Image editing with volume changes



Given image and audio



Given image and audio







Extension with Diffusion model (IJCV 2026)



CVPR LONG BEACH CALIFORNIA June 16-20, 2019

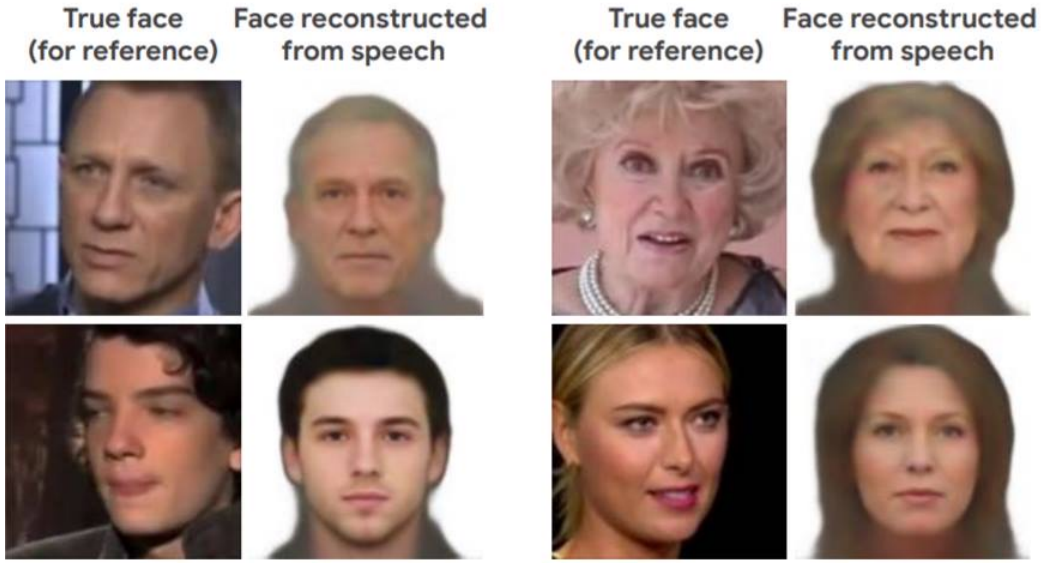
Speech2Face: Learning the Face Behind a Voice

Tae-Hyun Oh* Tali Dekel* Changil Kim* Inbar Mosseri,
William T. Freeman Michael Rubinstein Wojciech Matusik

MIT CSAIL Google Research
(* Equal contribution)

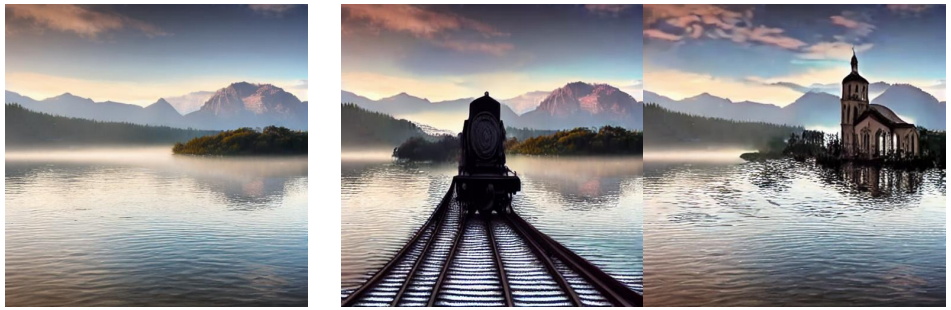
GIZMODO
and ≥ 19 int. news media





SoundBrush: Sound as a Brush for Visual Scene Editing

Kim Sung-Bin[1] Kim Jun-Seong[1] Junseok Ko[2]
Yewon Kim[1] Tae-Hyun Oh[1]
[1]POSTECH [2]Inha University

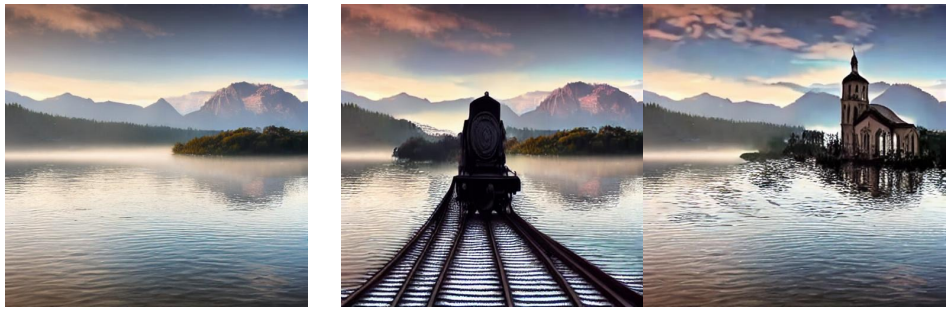




Code: https://soundbrush.github.io/

SoundBrush: Sound as a Brush for Visual Scene Editing

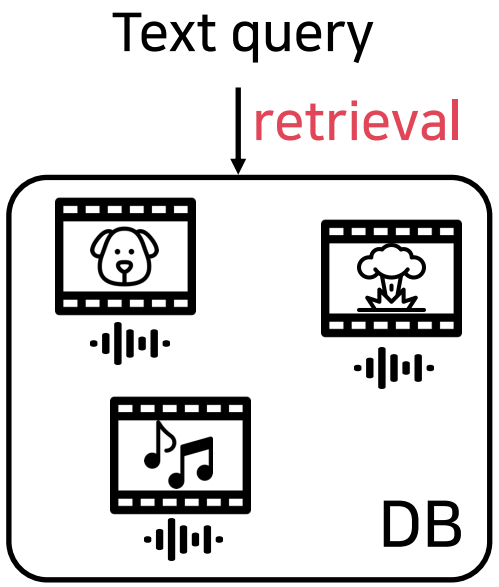
Kim Sung-Bin[1] Kim Jun-Seong[1] Junseok Ko[2]
Yewon Kim[1] Tae-Hyun Oh[1]
[1]POSTECH [2]Inha University



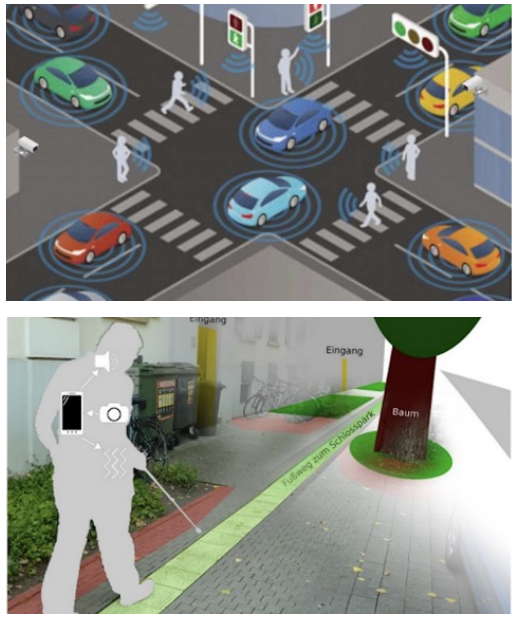


Code: https://soundbrush.github.io/

Applications



Sound file navigation



Applications for auditory impaired



Audio-guided media creation and media editing

Physical Understanding in Audio-Visual Learning

CVPR
JUNE 3-7, 2026
DENVER
COLORADO

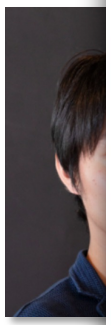
PAVAS: Physics-Av
Synt



Oh Hyun-Bin[1†]



Yuhta Takida[2]



Toshimit

[1]POSTECH

[2]Sony AI



† Work done during an internship at **Sony AI**

Explicit Physical Signal as Prior Information

Sound makes visual events feel real



Original Silent Film + Generating Audio

Foley: the craft behind realistic sound

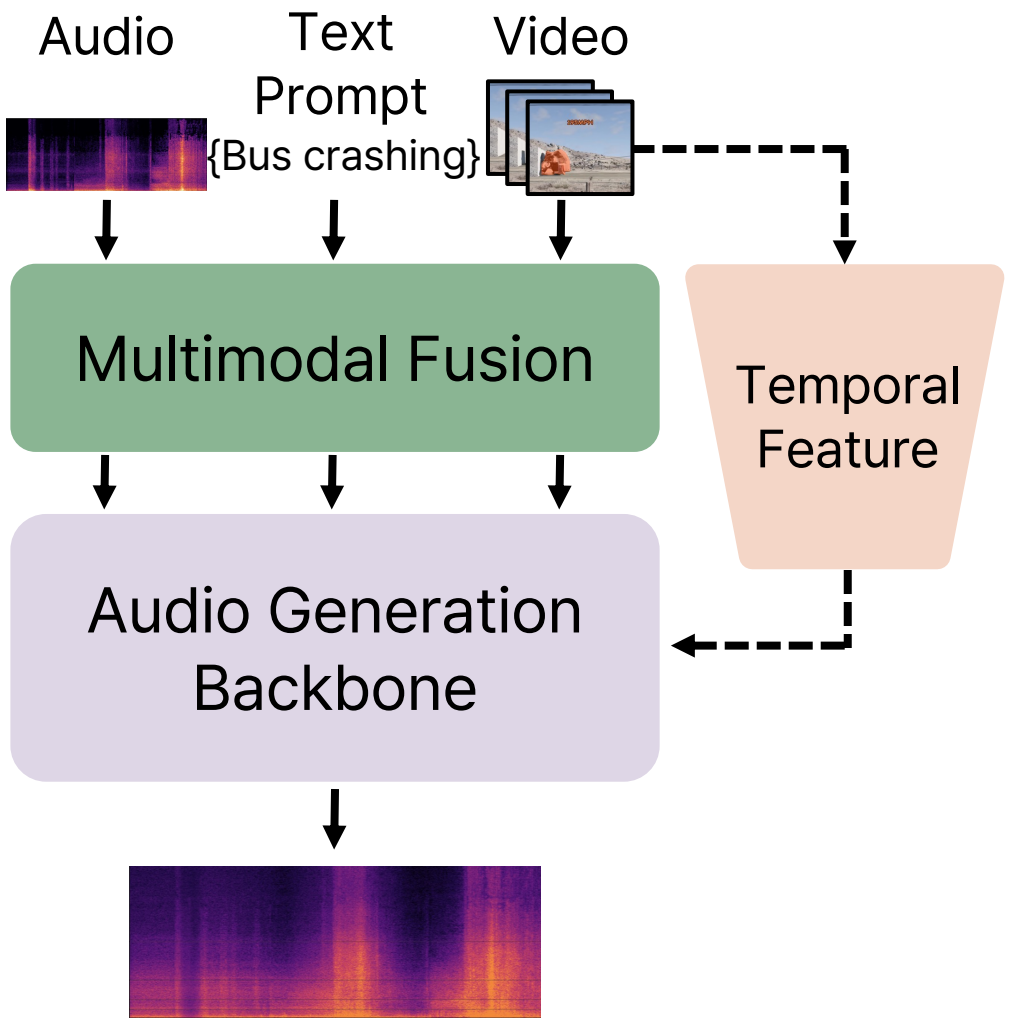
High-quality foley is labor-intensive and hard to scale

Video-to-audio (V2A) generation
= Scaling Foley sound from visual input by automation

What makes a good V2A model?

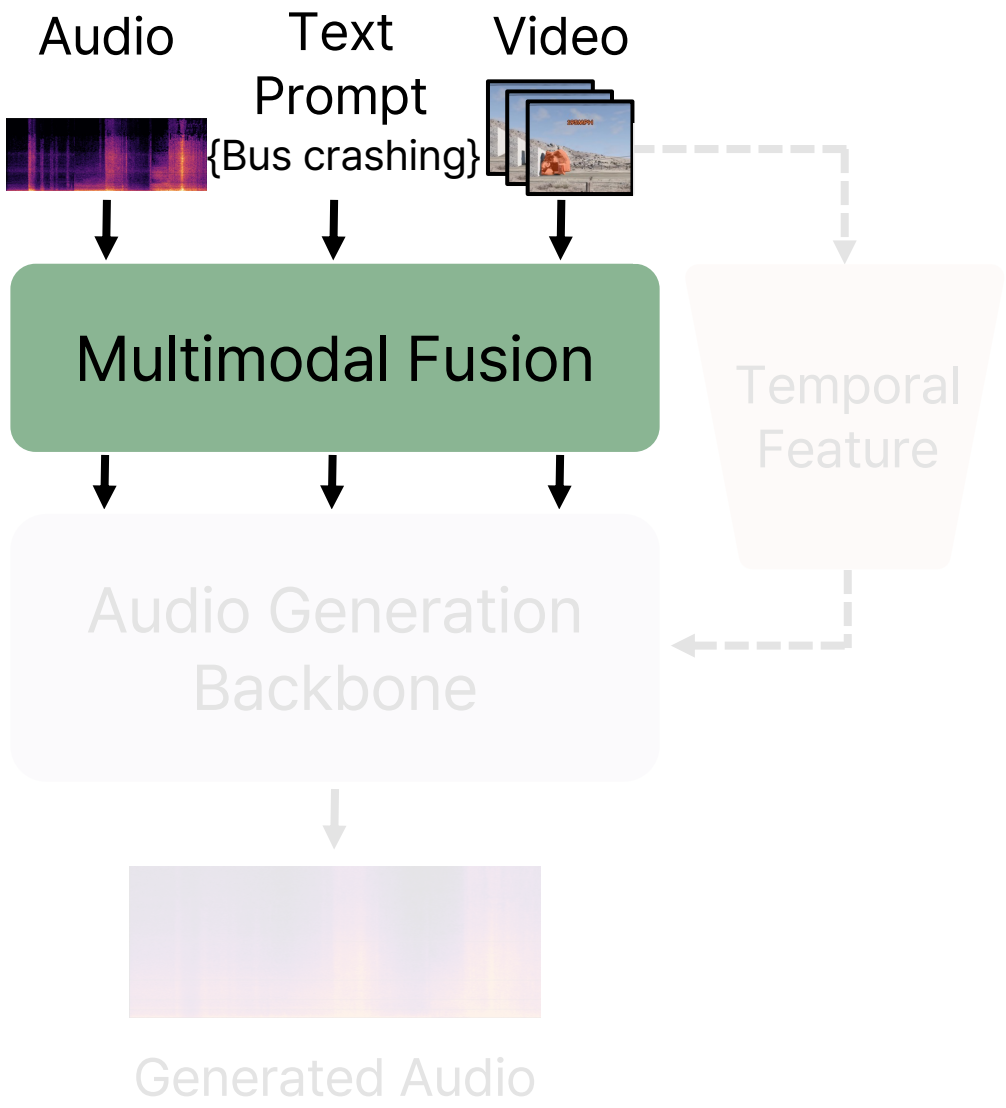
- What happens → Semantic alignment
- When it happens → Temporal alignment
- How realistic → Audio fidelity
- How objects move and touch → Physical consistency

Prior V2As improve semantics, timing, and fidelity



<Typical V2A pipeline>

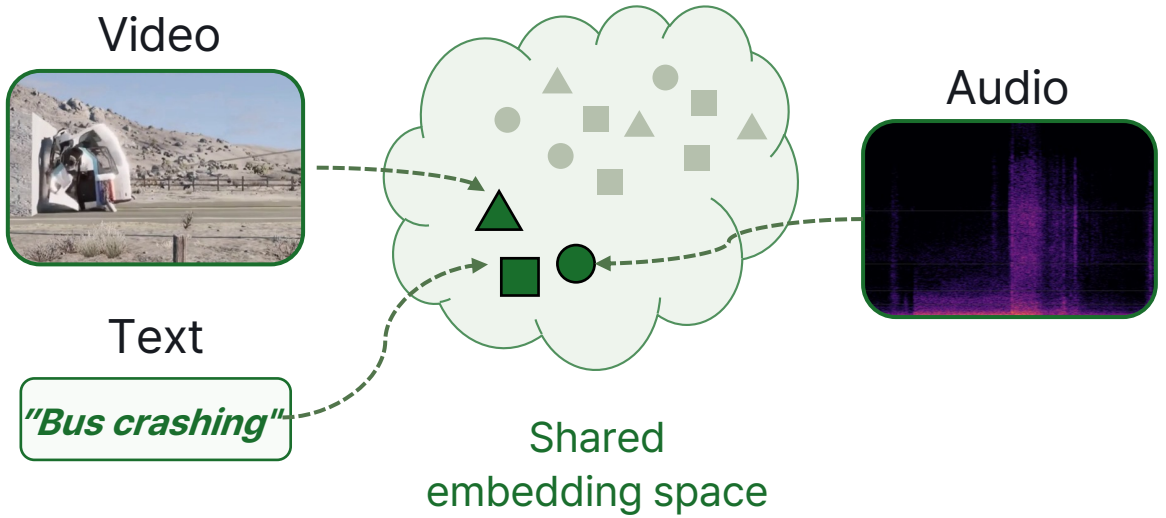
Prior V2As improve semantics, timing, and fidelity



<Typical V2A pipeline>

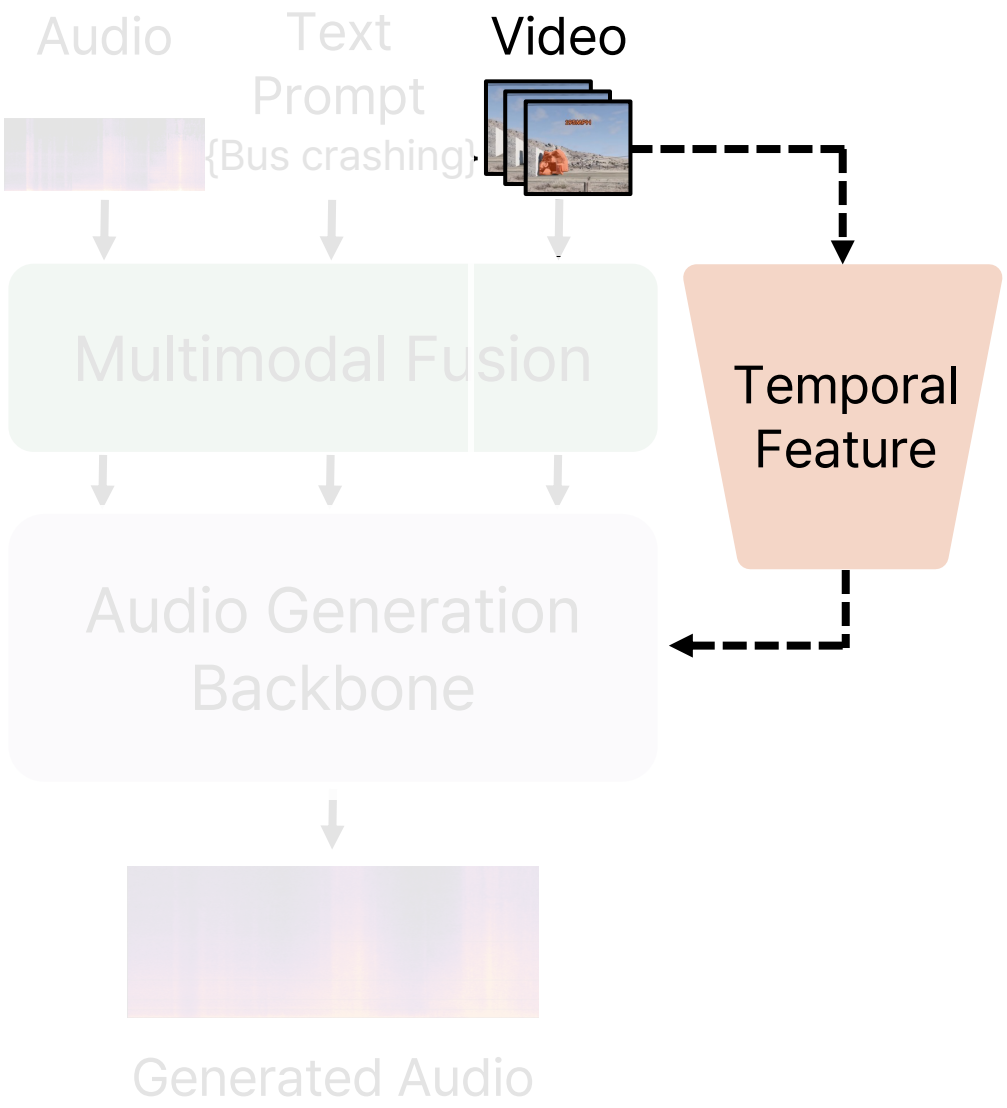
Semantic alignment

[DIFF-FOLEY (NeurIPS'23), V2A-Mapper (AAAI'24)]



What sound?

Prior V2As improve semantics, timing, and fidelity



<Typical V2A pipeline>

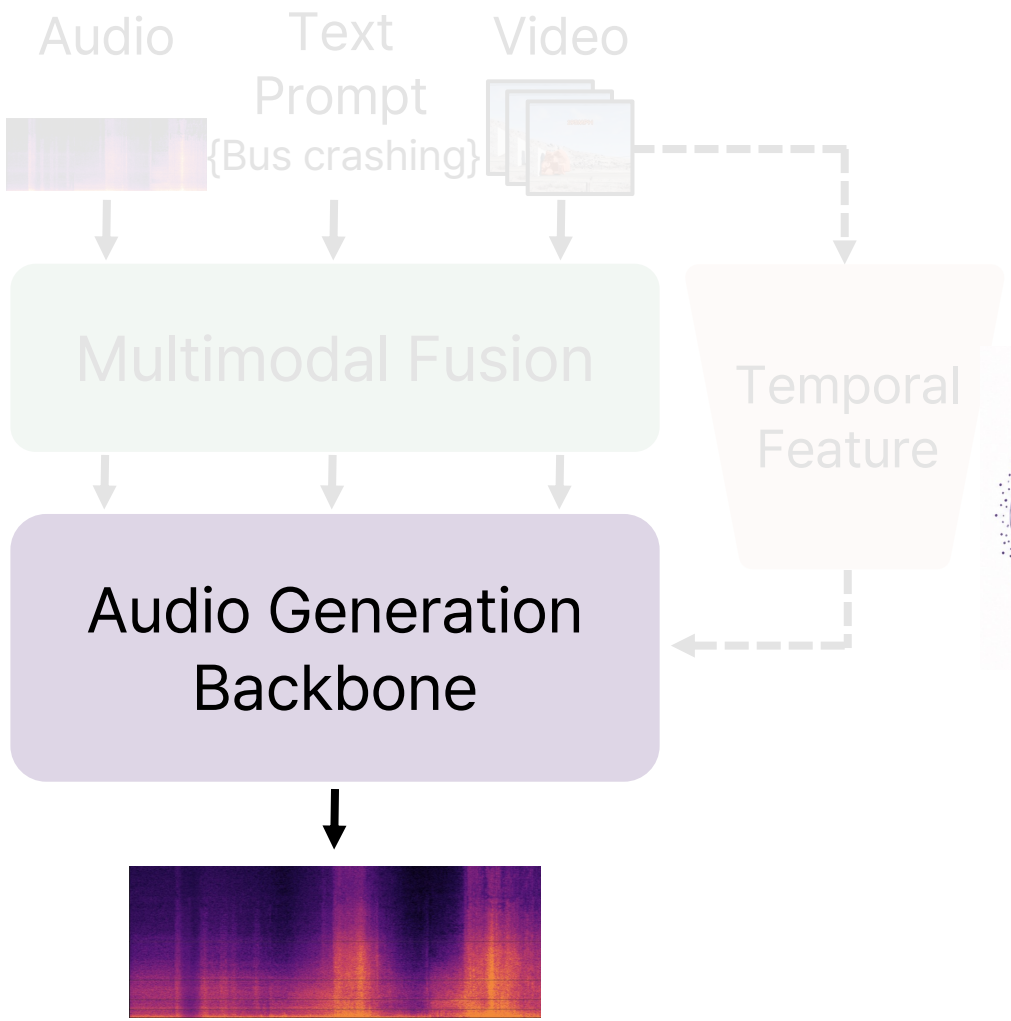
Temporal alignment

[ReWaS (AAAI'25), TARO (ICCV'25)]



When?

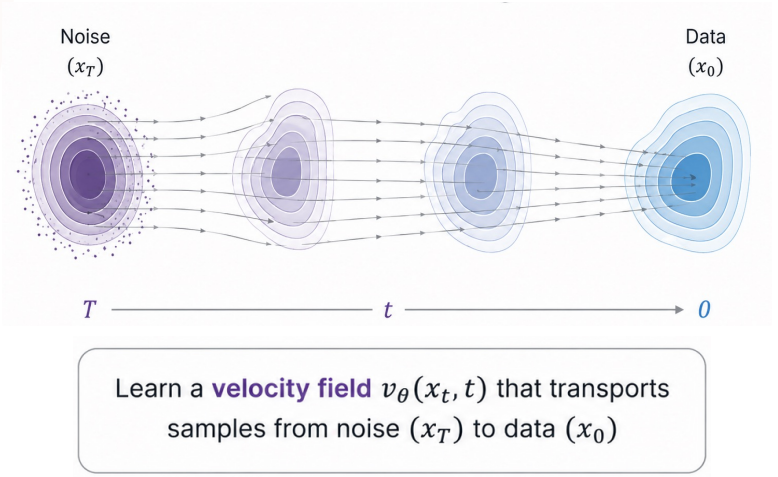
Prior V2As improve semantics, timing, and fidelity



<Typical V2A pipeline>

Audio fidelity

[Frieren (NeurIPS'24), MMAudio (CVPR'25)]



Flow matching



Combining heterogeneous dataset

How high-quality?

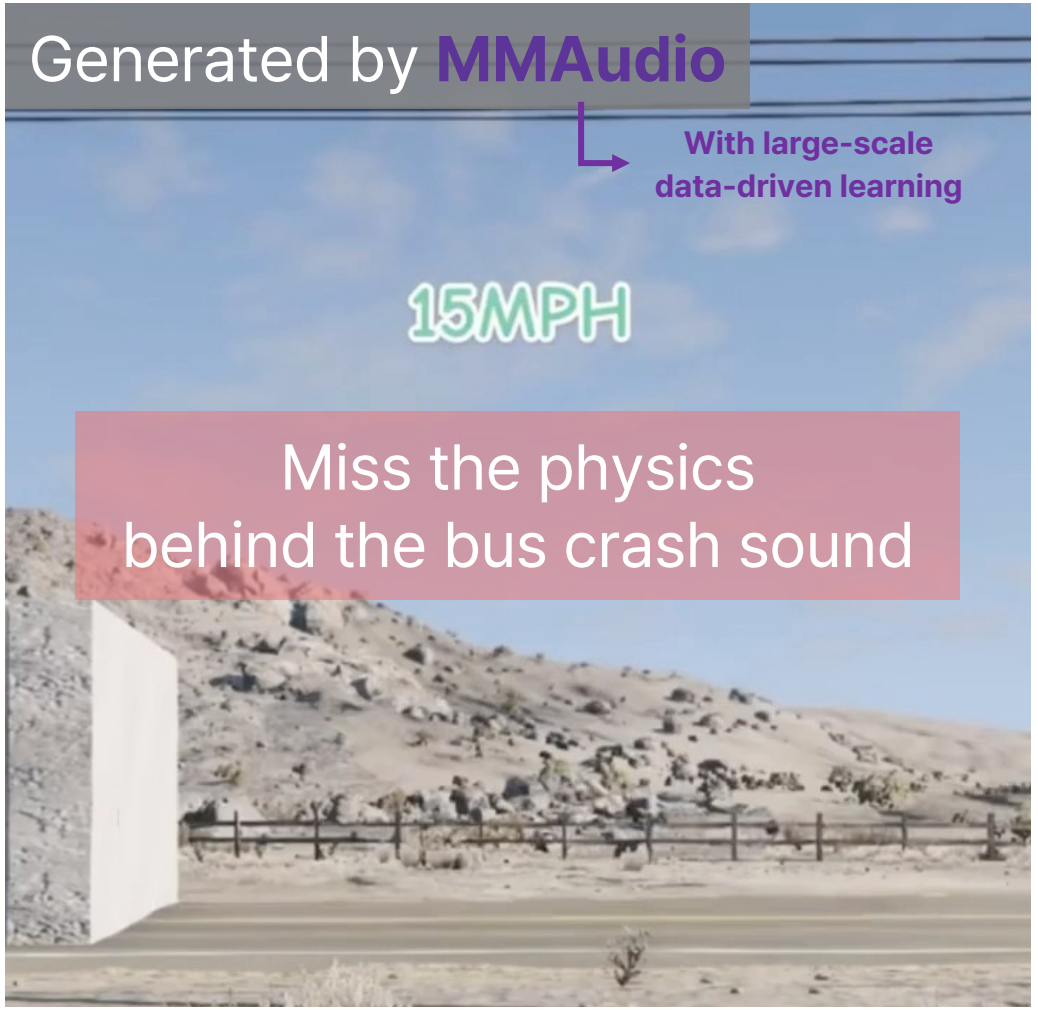
Imagine how this video should sound



Humans naturally form a physics-informed expectation of sound[1]

[1] Stekelenburg and Vroomen, "Predictive coding of visual-auditory and motor-auditory events: An electrophysiological study," Brain Research, 2015.

Does prior V2A match our expectation?

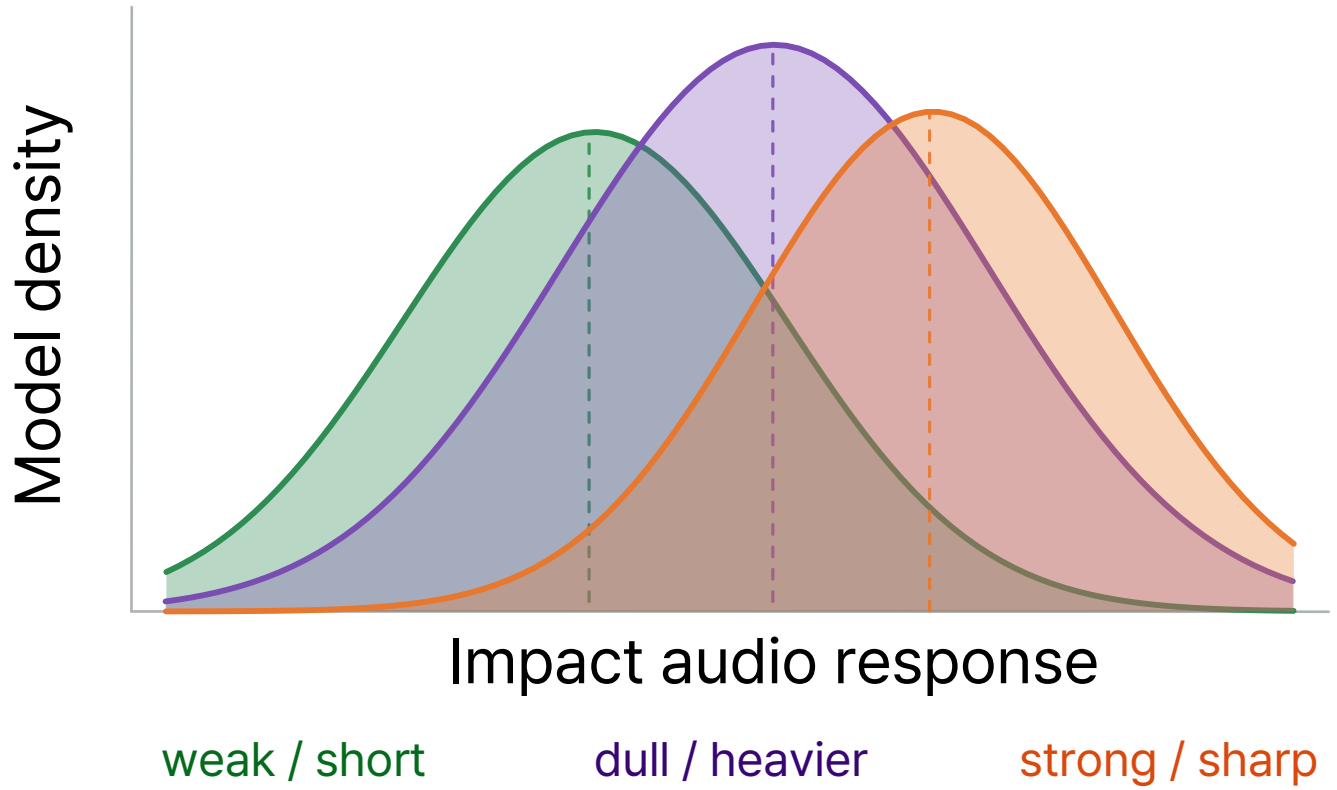


Data-driven co-occurrence learning is not enough to capture physical sound dynamics

Naïve co-occurrence leaves physical ambiguity

$$p_\theta(audio \mid video, text)$$

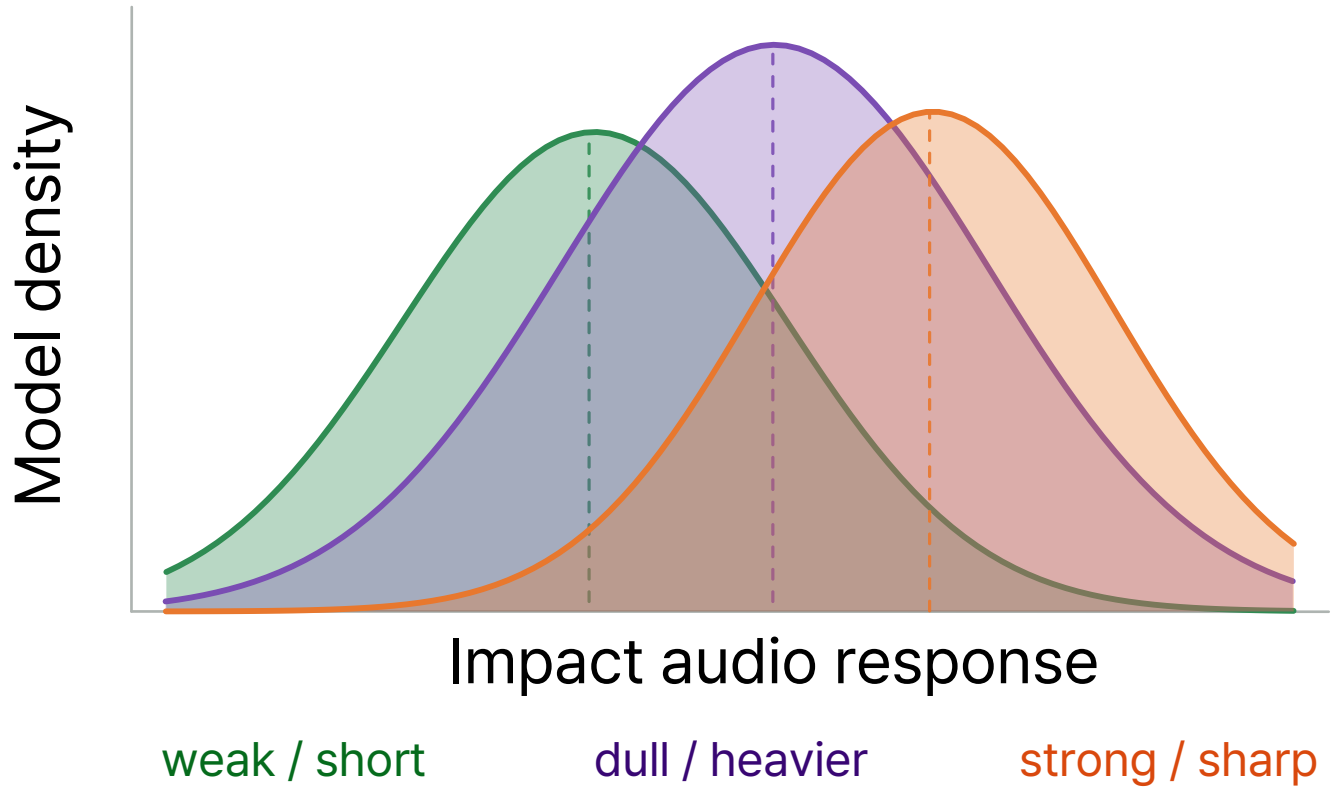
θ: learned parameters of the generative model



Physical condition disambiguates the audio mode

$$p_\theta(audio \mid video, text)$$

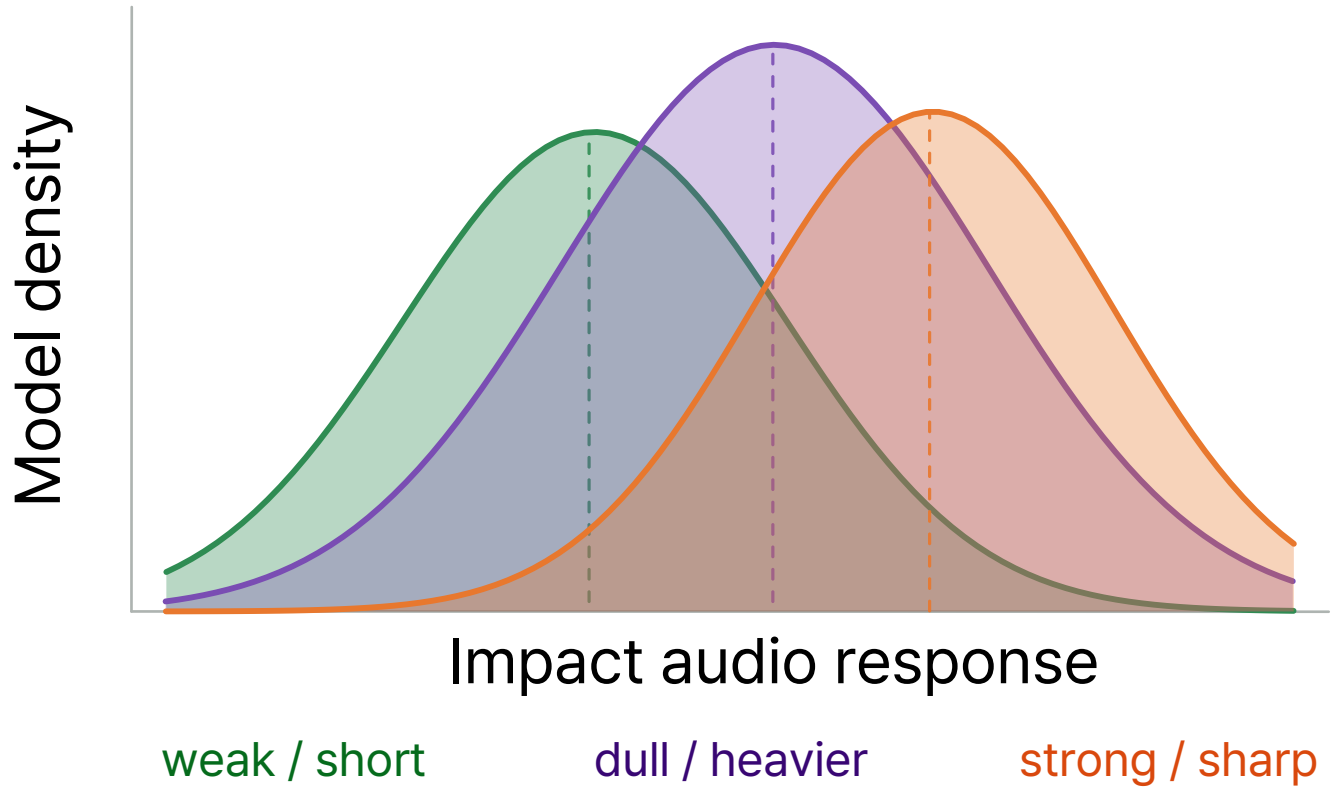
θ: learned parameters of the generative model



Physical condition disambiguates the audio mode

$$p_\theta(audio \mid video, text)$$

θ: learned parameters of the generative model

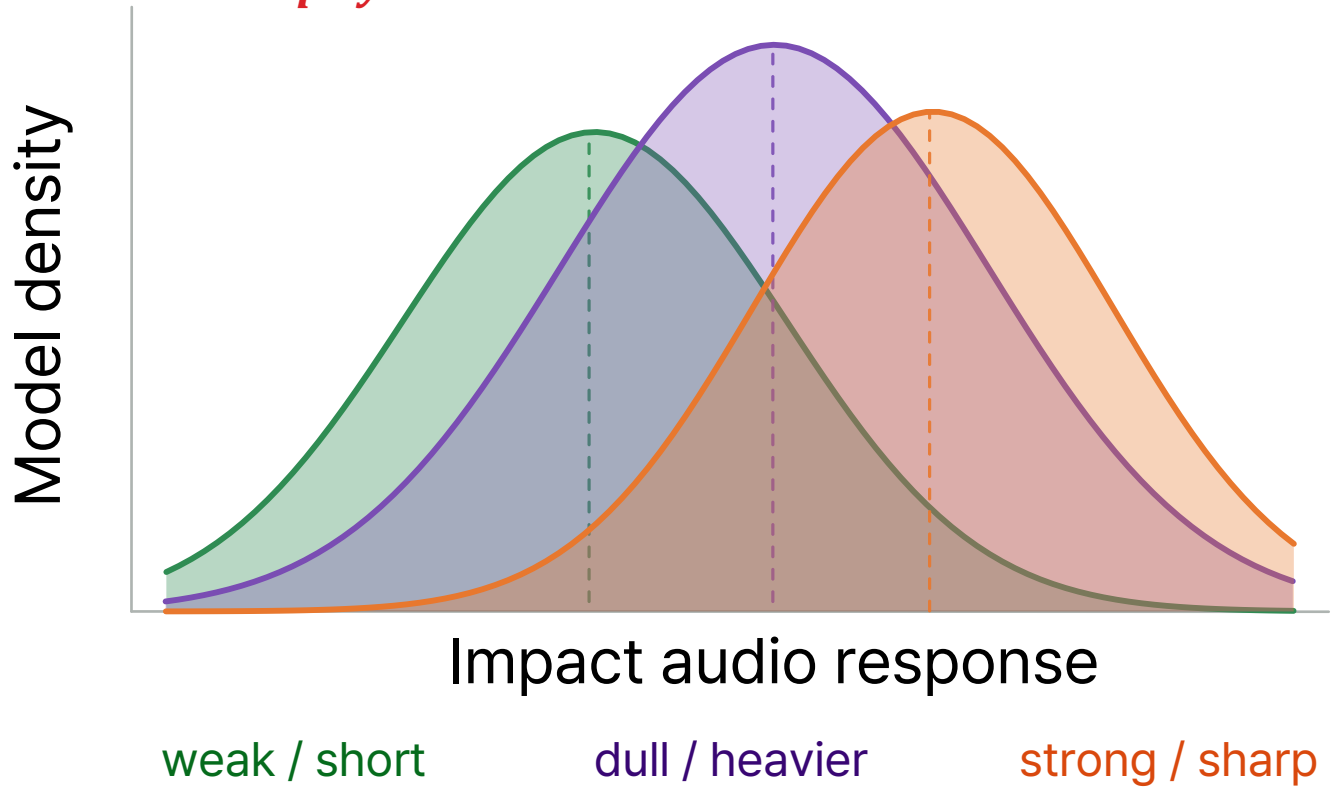


Physical condition disambiguates the audio mode

$$p_\theta(audio \mid video, text, z_{phys})$$

θ: learned parameters of the generative model

$$z_{phys} = \{object's\ mass, velocity\}$$

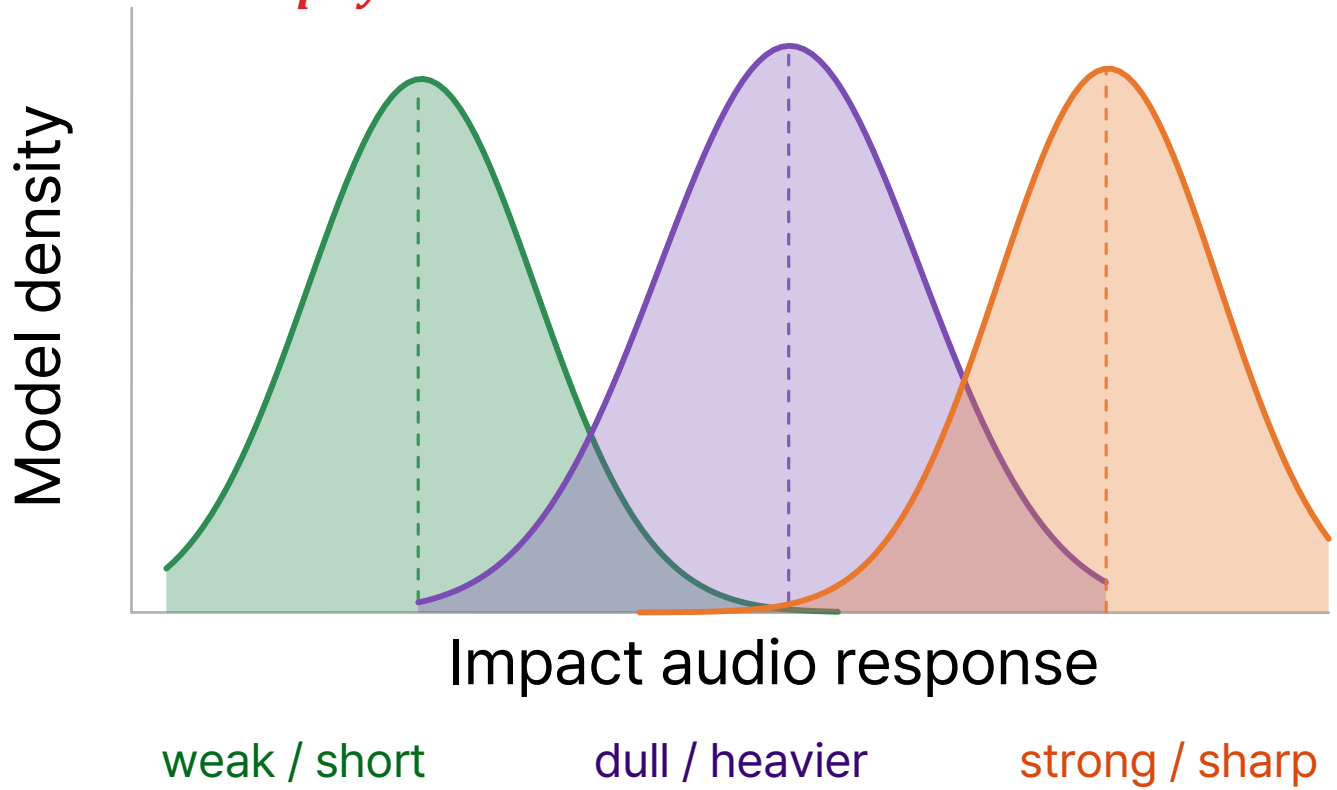


Physical condition disambiguates the audio mode

$$p_\theta(audio \mid video, text, z_{phys})$$

θ: learned parameters of the generative model

$$z_{phys} = \{object's\ mass, velocity\}$$



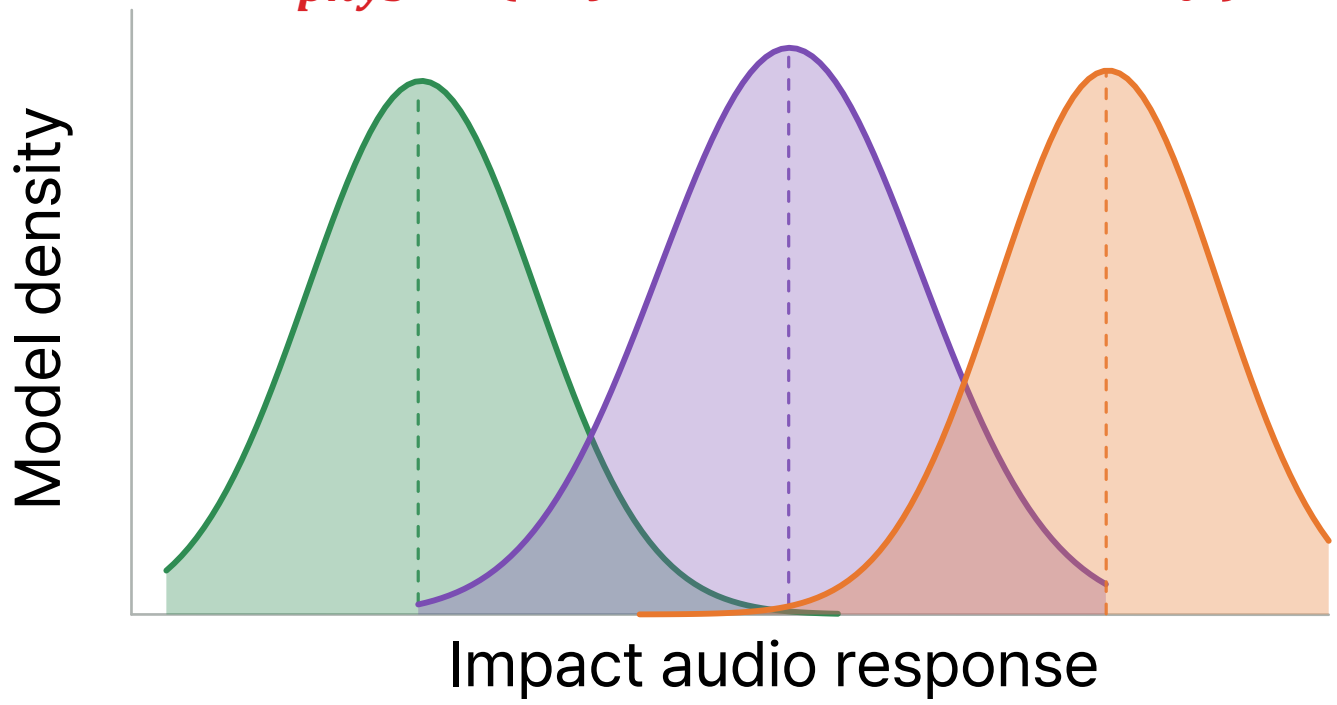
Note: Only when mutual information between true variable and conditions is large

Physical condition disambiguates the audio mode

$$p_\theta(audio \mid video, text, z_{phys})$$

θ: learned parameters of the generative model

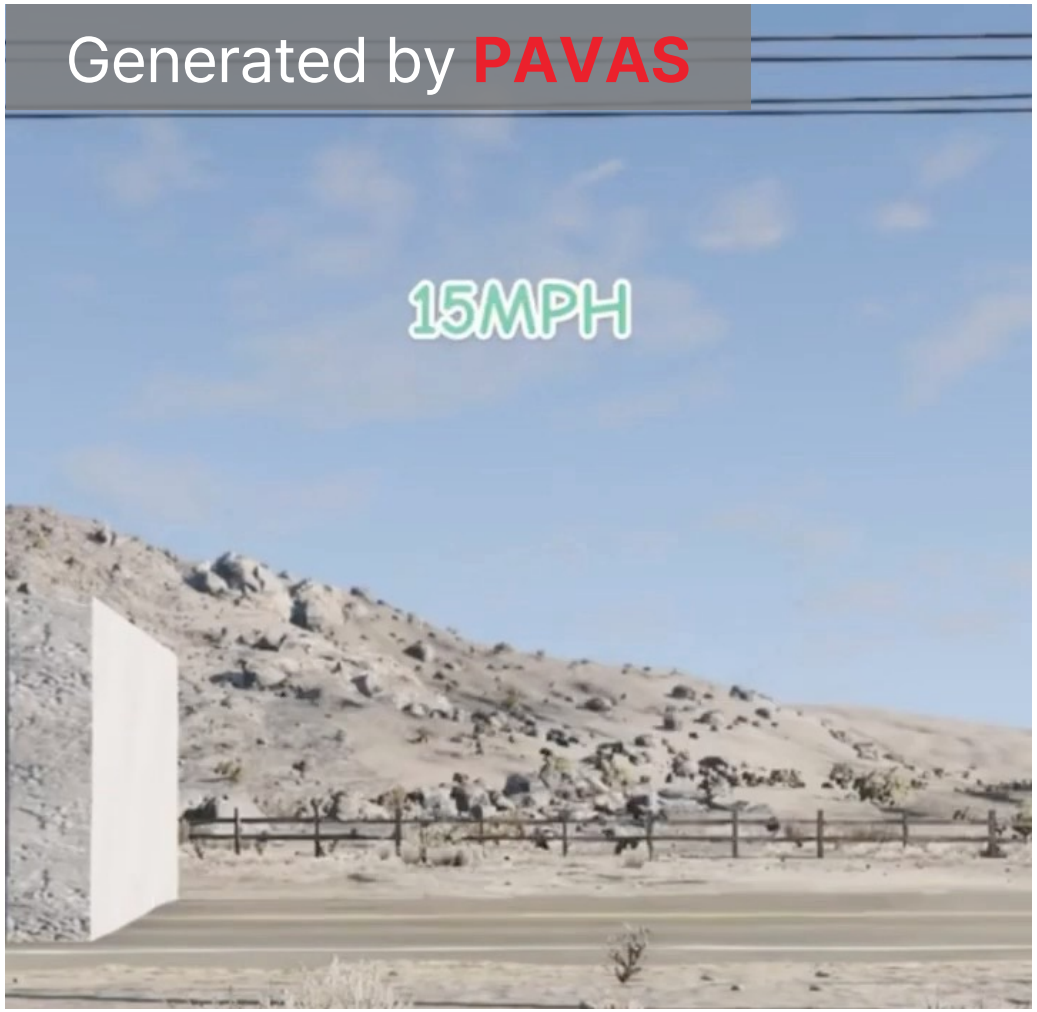
$$z_{phys} = \{object's\ mass, velocity\}$$



Explicit physical conditions reduce ambiguity in video-to-audio generation

Note: Only when mutual information between true variable and conditions is large

The missing variable: physical state



Physics Parameter Estimator (PPE)

Video → Object-level physical state z_{phys}

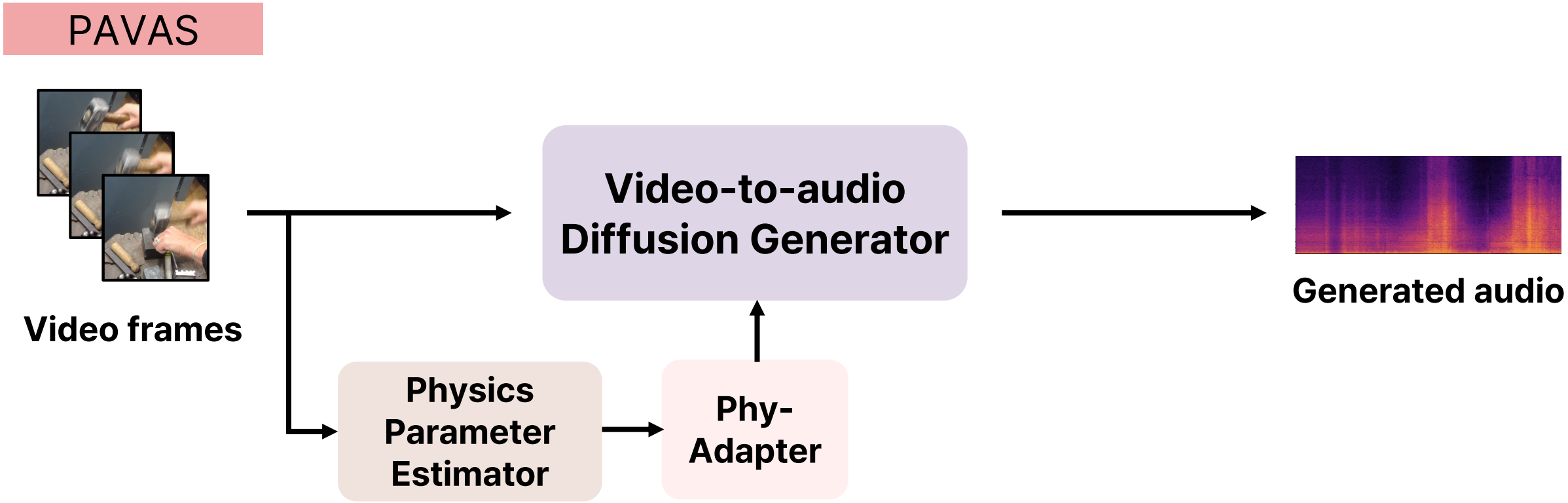
- Moving object localization

- Mass estimation

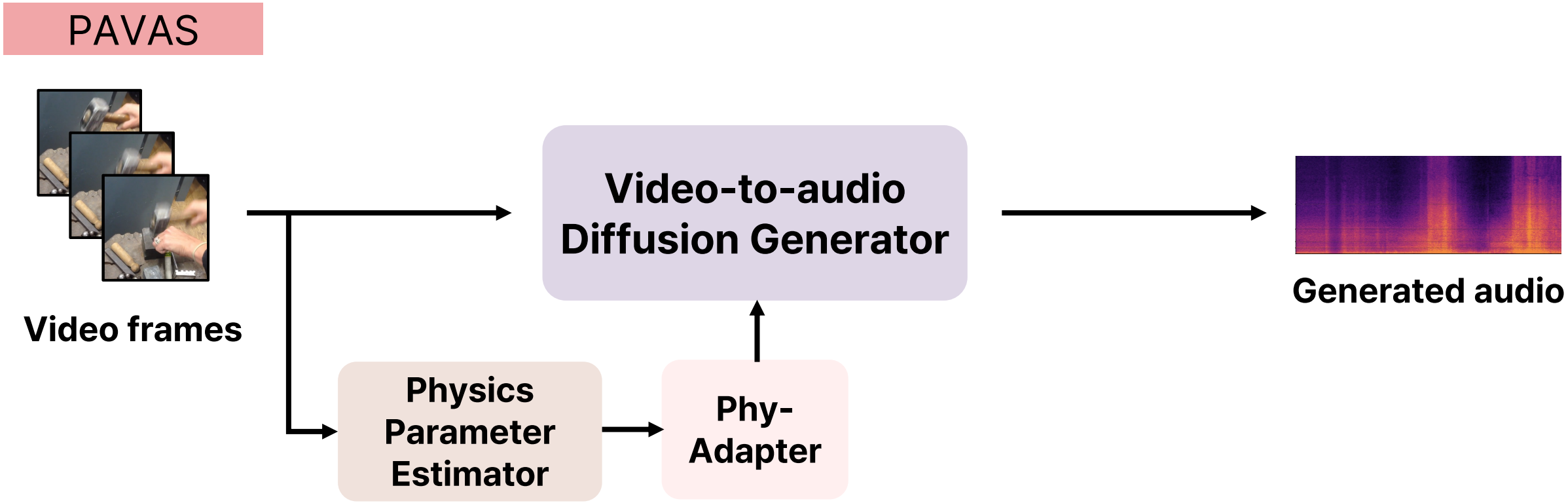
- Velocity estimation

The impact now sounds better follows the visible dynamics

PAVAS: Estimate physics, and generate audio



PAVAS: Estimate physics, and generate audio

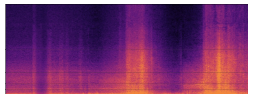


PPE turns video into object-level physics cues



Video frames

Video-to-audio Diffusion Generator



Generated audio

Physics Parameter Estimator

Phy-Adapter

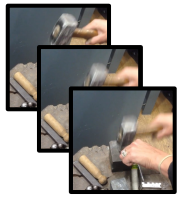


Object: Hammer

Mass: 3.5kg

Object mass

VLM



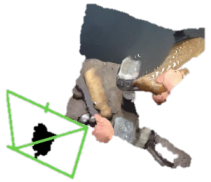
Video frames

Segment Anything



Object masks

Dynamic 3D Recon.



3D Point clouds

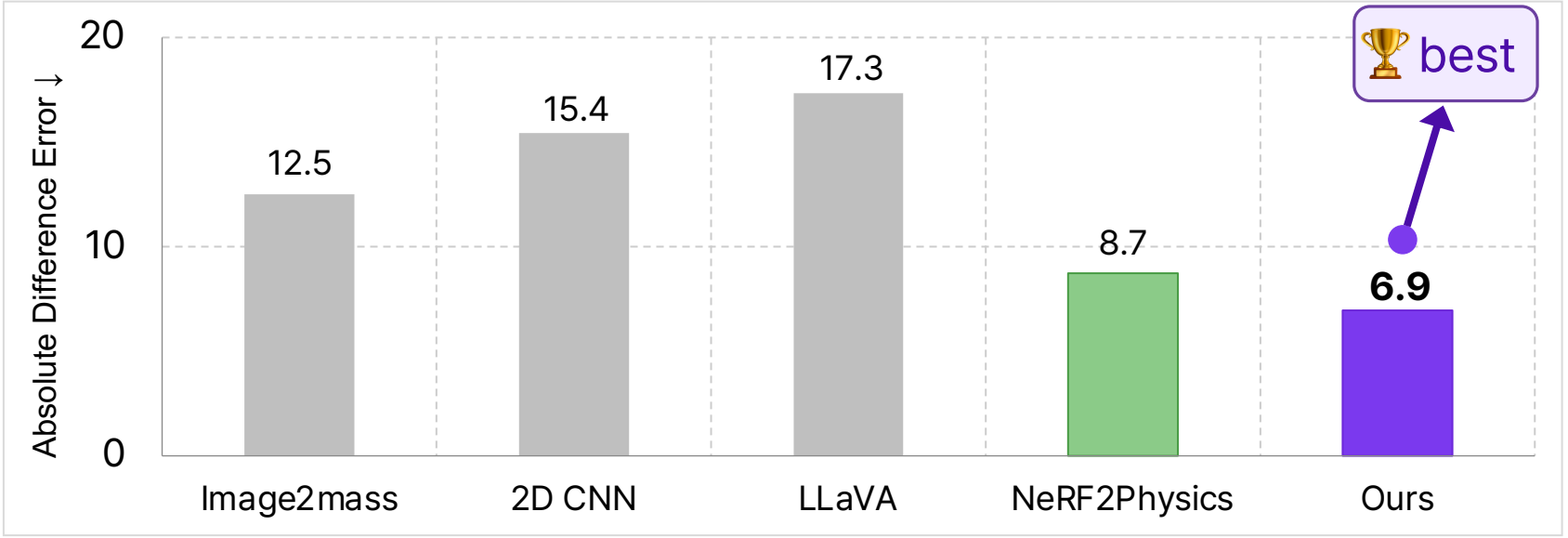
Inverse Proj.

Object velocity

PPE is a surprisingly strong physics estimator

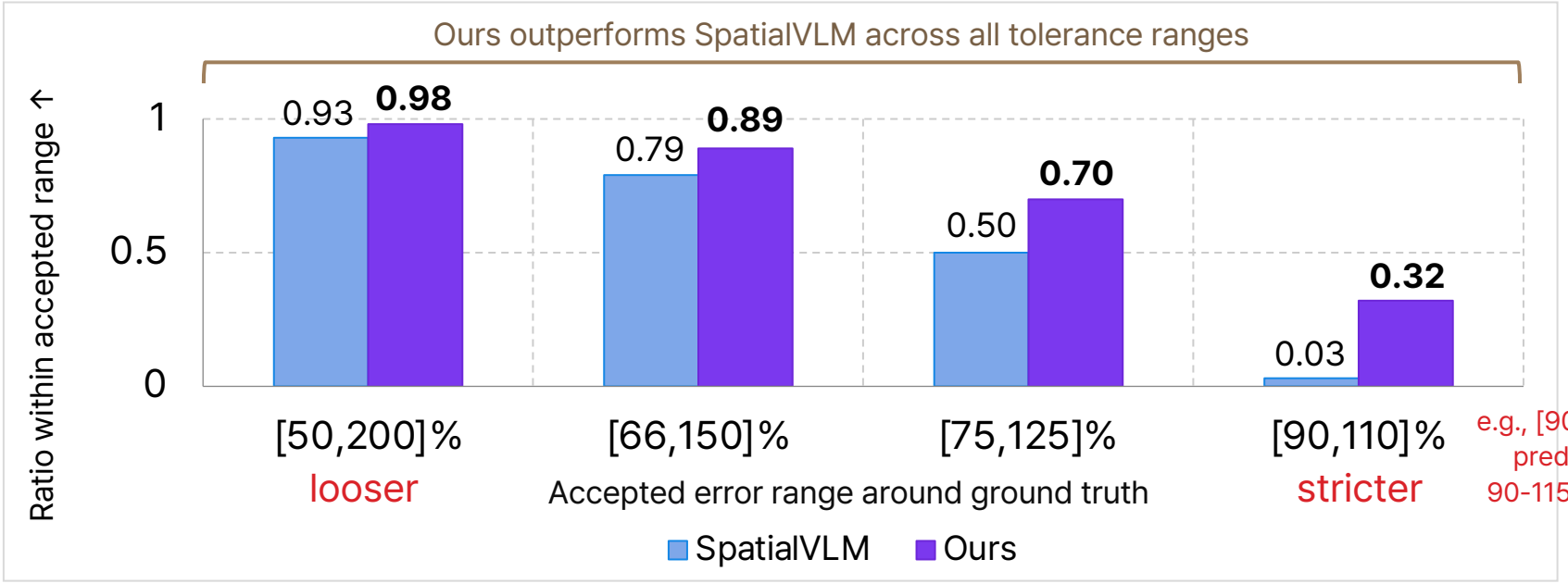
Mass estimation Performance

Lower is better



Velocity estimation Performance

Higher is better

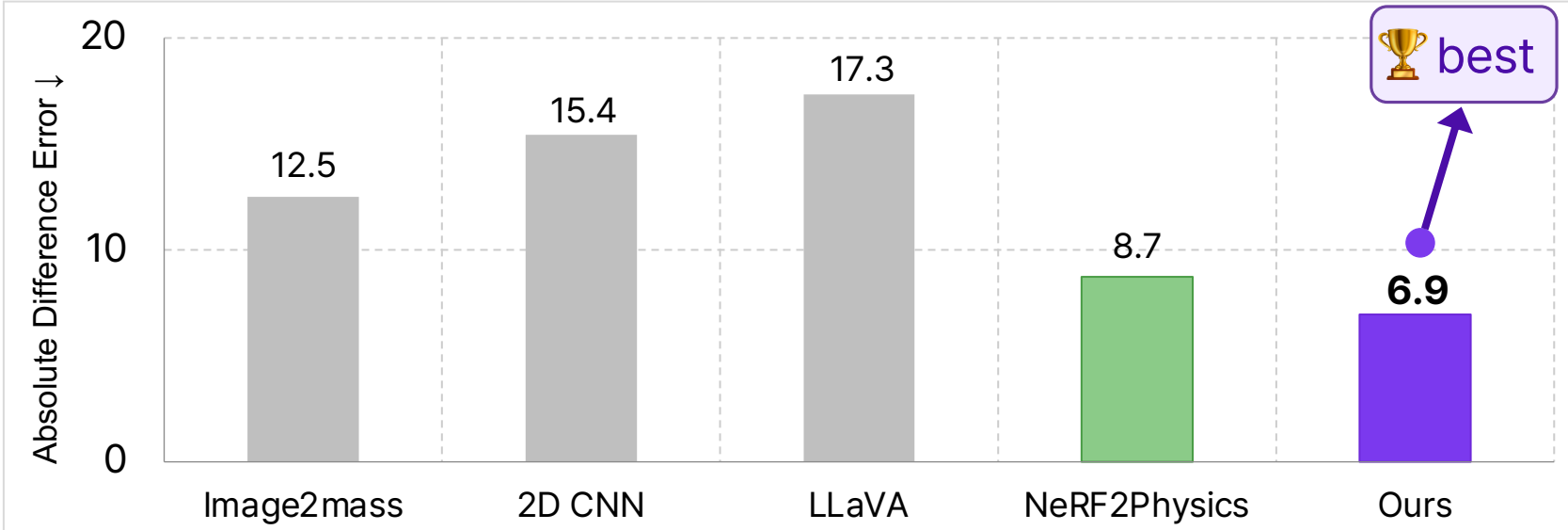


e.g., [90, 110]% means prediction within 90-115% of GT value

PPE is a surprisingly strong physics estimator

Mass estimation Performance

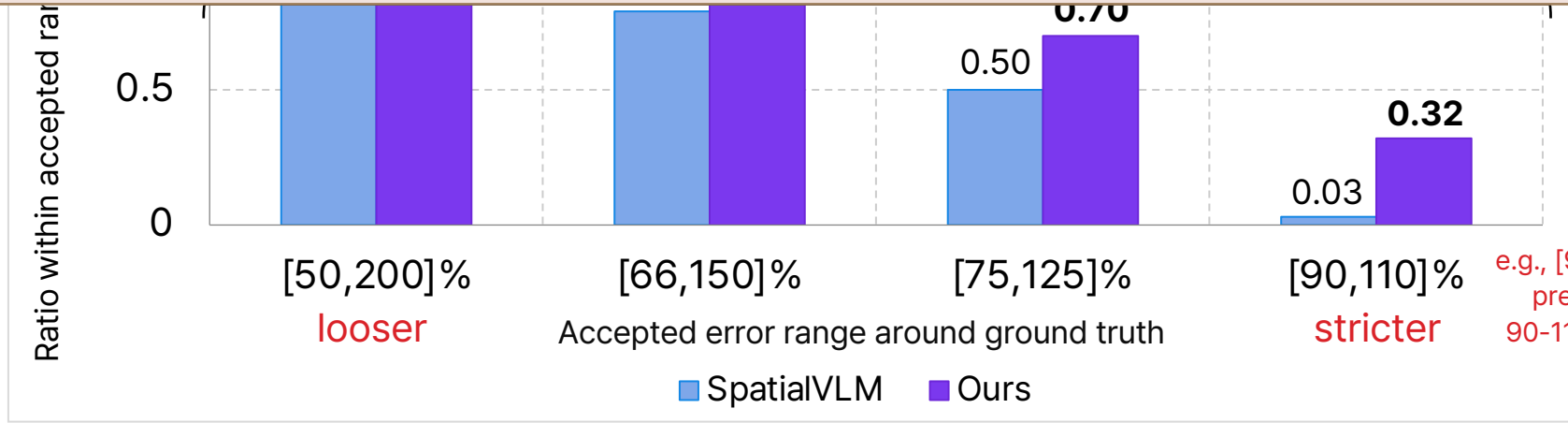
Lower is better



PPE estimates more reliable physical values than existing expert models for physics estimation

Velocity estimation Performance

Higher is better



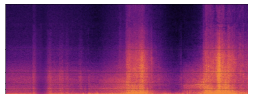
e.g., [90, 110]% means prediction within 90-115% of GT value

Phy-Adapter turns physics into control signals



Video frames

Video-to-audio Diffusion Generator



Generated audio

Physics Parameter Estimator

Phy-Adapter



Object: Hammer

Mass: 3.5kg

Object mass

VLM



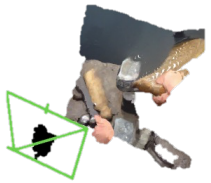
Video frames

Segment Anything



Object masks

Dynamic 3D Recon.

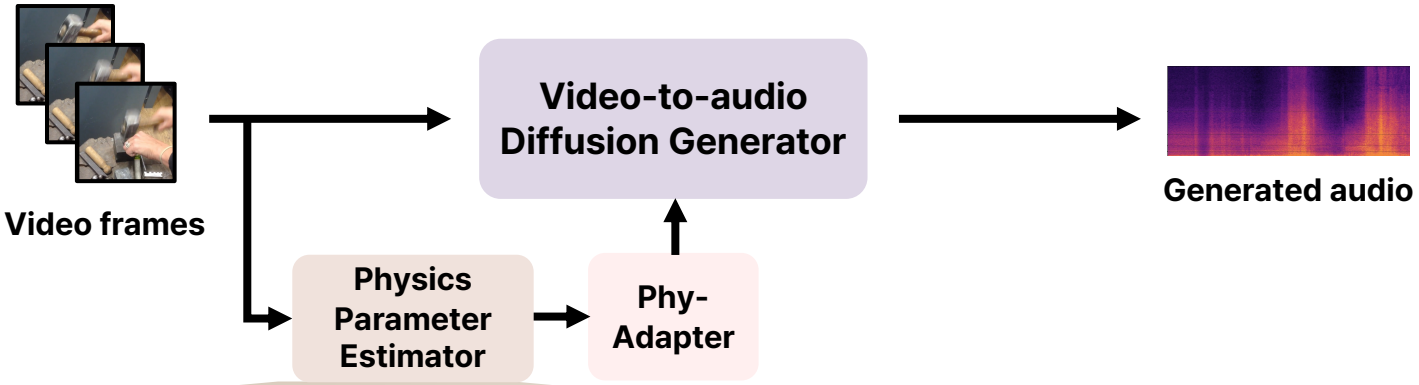


3D Point clouds

Inverse Proj.

Object velocity

Phy-Adapter turns physics into control signals



Object mass

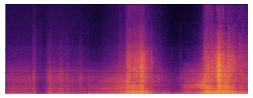
Object velocity

Phy-Adapter turns physics into control signals



Video frames

Video-to-audio Diffusion Generator



Generated audio

Physics Parameter Estimator

Phy-Adapter

mass

kg 3.5 kg

Object mass

MLP

...

Mass embedding

v(t) velocity

t

Object velocity

MLP

...

Velocity embedding

Diffusion Transformer Blocks

Block l

Base feature h

Physics feature p

Learnable gate α

Delta modulation Δ

Block l+1

Base feature h

Physics feature p

Learnable gate α

Delta modulation Δ

Block l+2

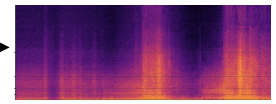
Base feature h

Physics feature p

Learnable gate α

Delta modulation Δ

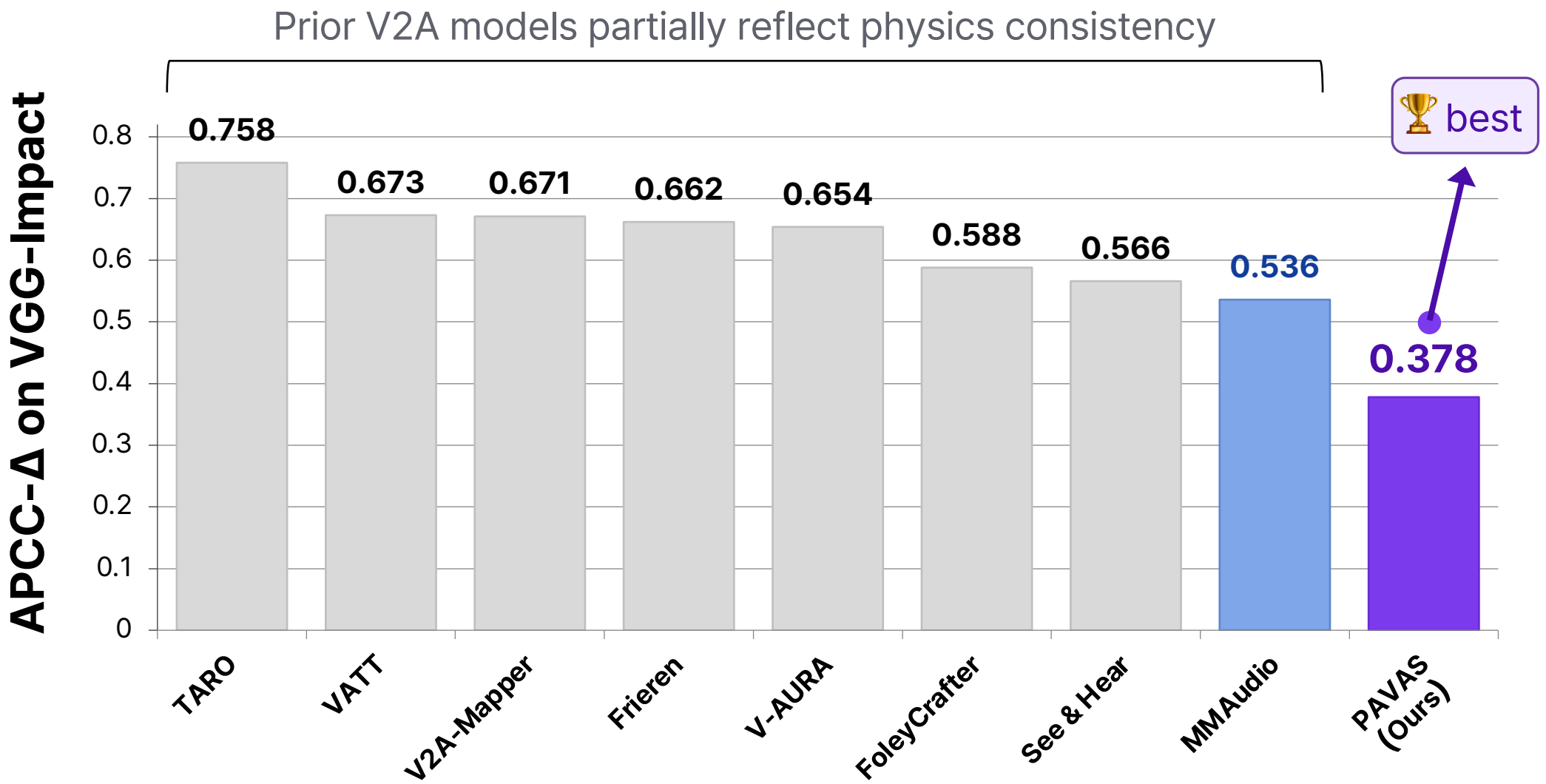
...



Generated audio

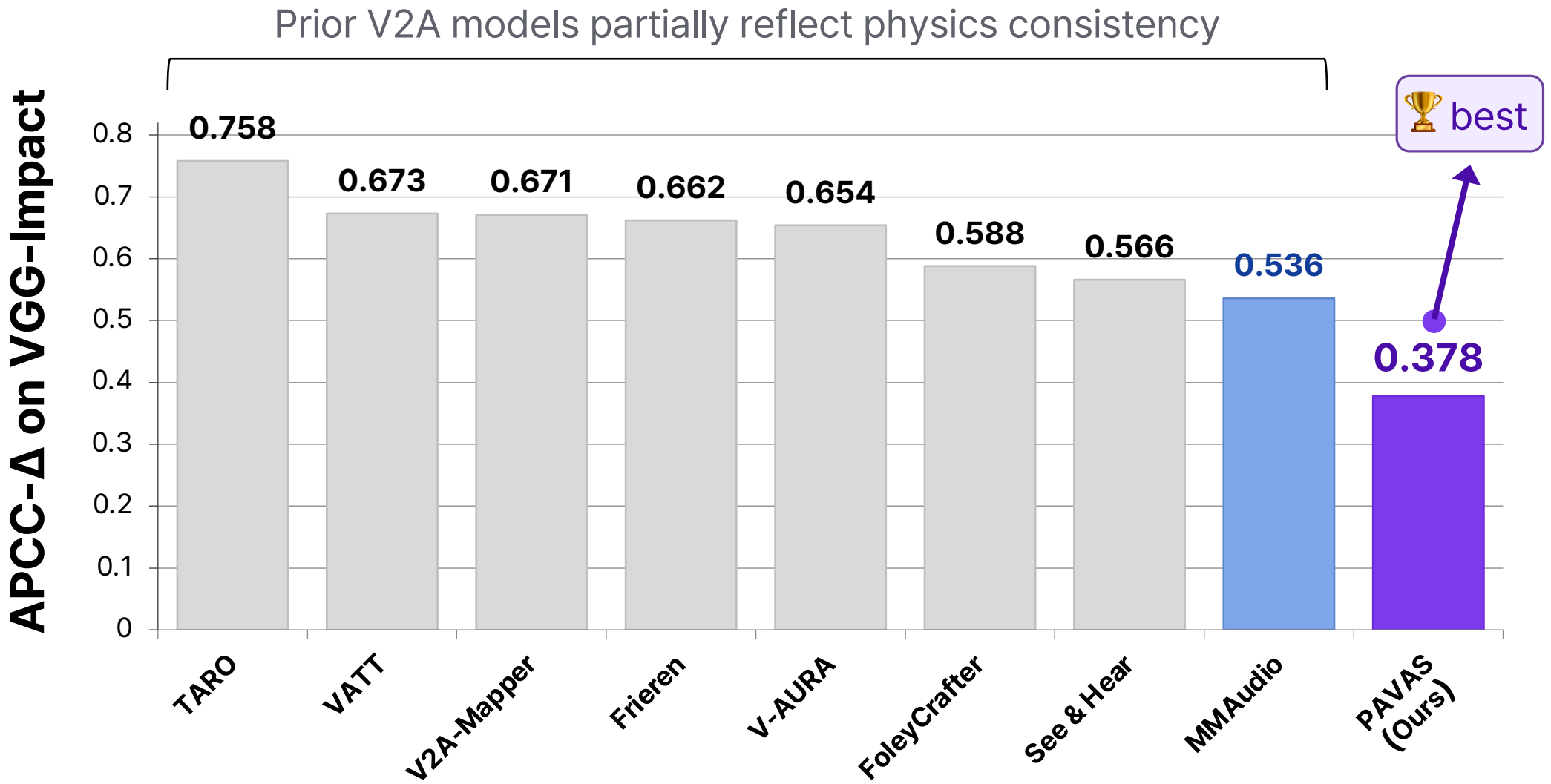
Modulated feature h′ = Base feature h + Delta modulation Δ

How well do prior models reflect physics?



How well do prior models reflect physics?

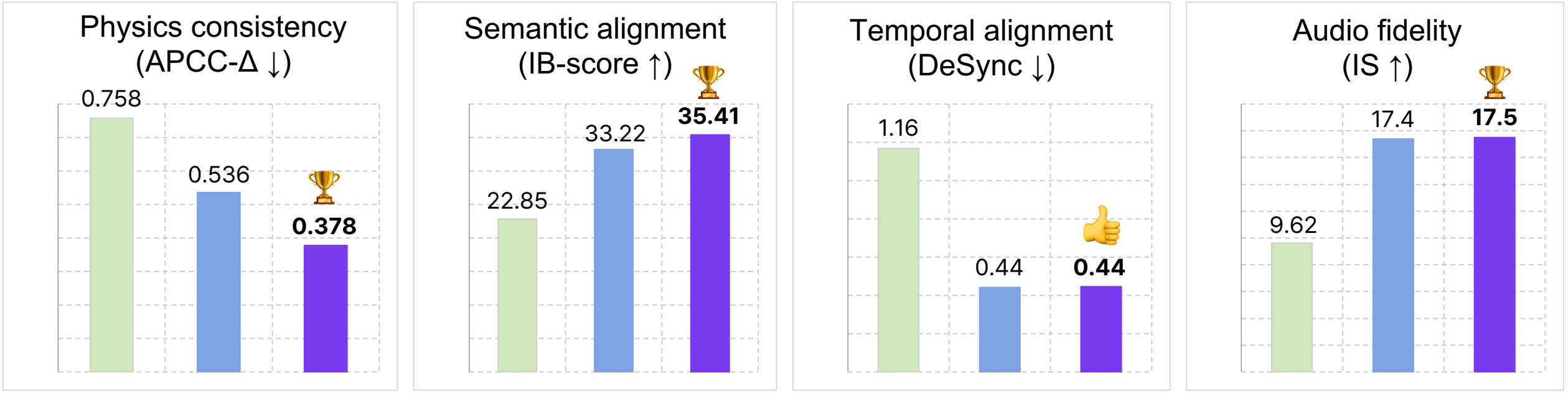
PAVAS shows the lowest APCC-Δ, indicating the best physics consistency



PAVAS improves more than physics consistency

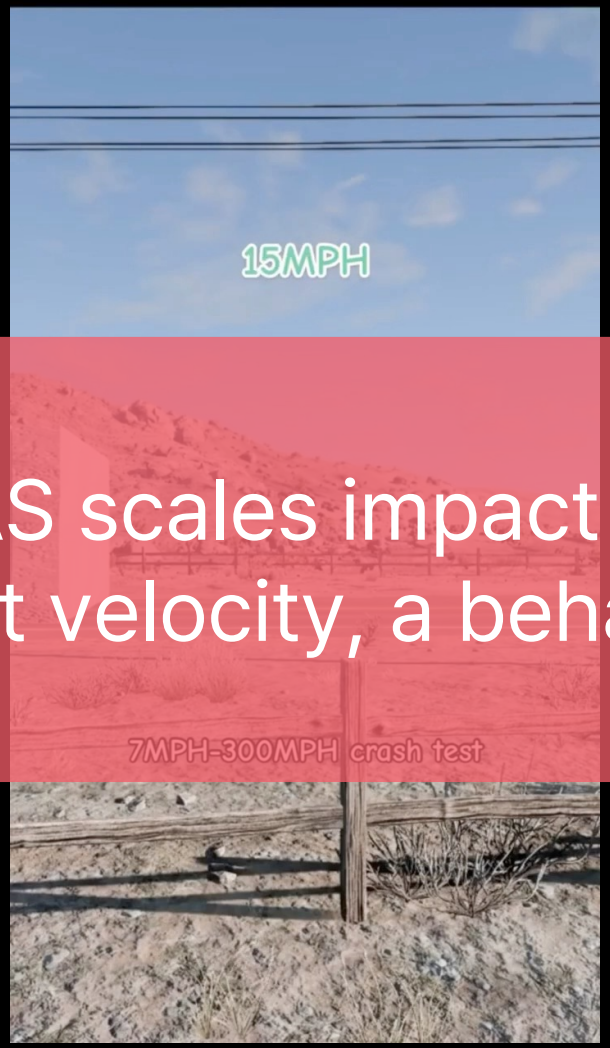
TARO [ICCV'25] | MMAudio [CVPR'25] | PAVAS (Ours) [CVPR'26]

On VGG-Impact dataset



Explicit physics cues improve not only physical consistency, but also semantic and temporal alignments, and audio fidelity.

Velocity change



MMAudio-L
[CVPR 2025]



PAVAS (Ours)
[CVPR 2026]

PAVAS scales impact sharpness and loudness with object velocity, a behavior missing in MMAudio.

Mass change



MMAudio-L
[CVPR 2025]



PAVAS (Ours)
[CVPR 2026]

PAVAS shows clearer loudness and sharpness in response to mass change.

Semantic (material) change



MMAudio-L
[CVPR 2025]



PAVAS (Ours)
[CVPR 2026]

Semantic (material) change



MMAudio-L
[CVPR 2025]



PAVAS (Ours)
[CVPR 2026]

Semantic (material) change



MMAudio-L
[CVPR 2025]



PAVAS (Ours)
[CVPR 2026]

PAVAS adapts impact sounds to material changes, whereas MMAudio often ignores them.

Qualitative comparison on VGGSound



TARO
[ICCV 2025]



MMAudio-L
[CVPR 2025]



PAVAS (Ours)
[CVPR 2026]

Qualitative comparison on Real-World Video

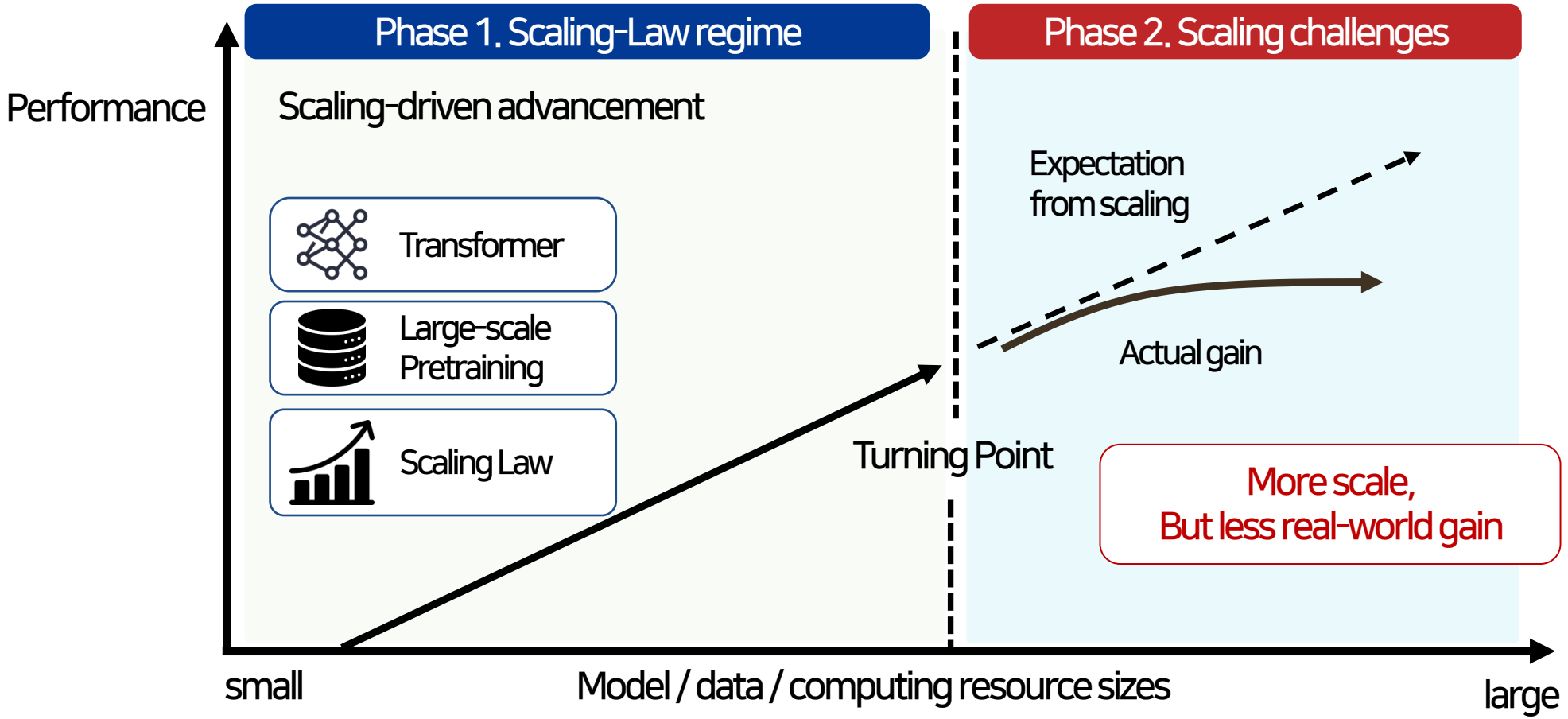


MMAudio-L
[CVPR 2025]



PAVAS (Ours)
[CVPR 2026]

Conclusion


Phase 1. Scaling-Law regime
Phase 2. Scaling challenges
Performance
Scaling-driven advancement
Transformer
Large-scale Pretraining
Scaling Law
Expectation from scaling
Actual gain
Turning Point
More scale,
But less real-world gain
small
Model / data / computing resource sizes
large

Conclusion

- Large-scale data-driven training alone is not sufficient.
  - We need more inductive bias
- The ways to inject prior
  - Supervised data (Sound source localization)
  - Pre-trained model / curriculum learning (Sound-to-Visual event generation)
  - Explicit physical signal (Video-to-audio generation)
- From mere visual-sound co-occurrence to learning causal relationship

Generative Cross-modal Machine Perception

Until now

- Ambient sound -



Audio event



Visual event

Related direction

- Speech to talking head -



Input Speech



Conclusion

Past





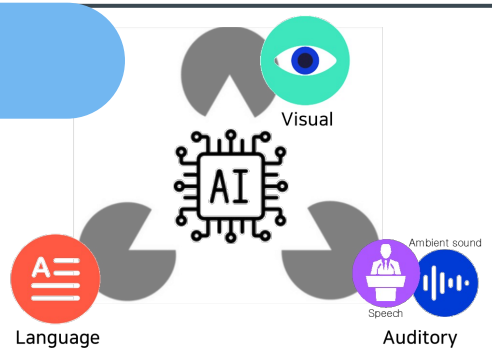
Common Cameras

capture the visible

Common AI

understand the visible

Future



Generative multi-modal perception

capture and understand the visible & the invisible

Thank you for listening



Thanks to coauthors

