

# IMAGEBIND: One Embedding Space To Bind Them All

Facebook AI Research (FAIR)

# Авторы



**Rohit Girdhar**  
PhD в Carnegie Mellon University

Internships at DeepMind, Adobe  
Research, Facebook AI



**Ishan Misra**  
PhD at the Robotics Institute at  
Carnegie Mellon University

SCS Distinguished  
Dissertation Award (Runner  
Up) 2018



**Mannat Singh**  
worked at Bloomberg and Goldman  
Sachs



**Armand Joulin**  
postdoc at Stanford University

## Предыдущие работы от авторов ImageBind

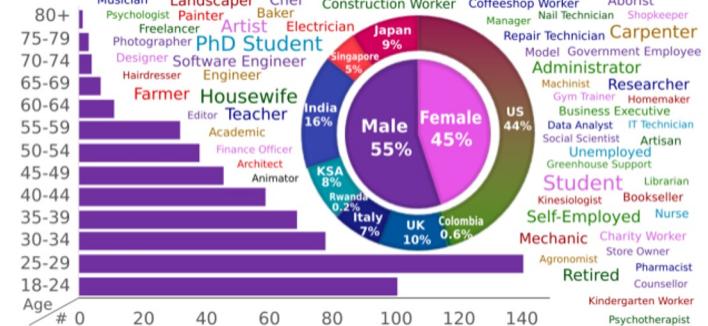
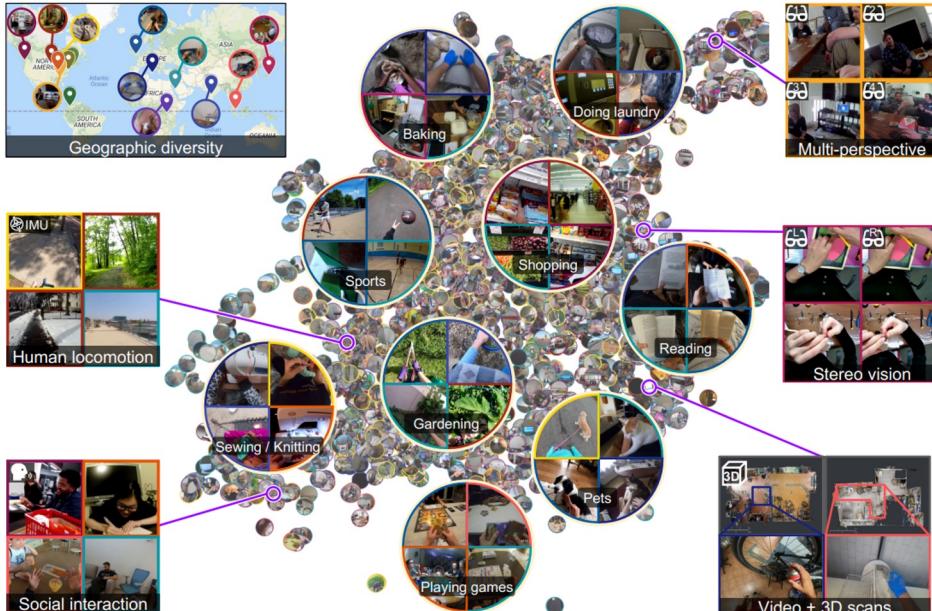
- Ego4D: Around the World in 3,000 Hours of Egocentric Video
- DETIC (Detecting Twenty-thousand Classes using Image-level Supervision)
- Omnivore: A Single Model for Many Visual Modalities
- OmniMAE: Single Model Masked Pretraining on Images and Videos

# Ego4D: Around the World in 3,000 Hours of Egocentric Video

- Ego4D - огромный датасет, состоящий из видео, описывающих различную деятельность человека
  - 3,025 часов видео
  - 855 уникальных носителей камер
  - 74 локации из 9 стран
- Представлены RGB-видео, аудио, 3D-мешы, детекция направления взгляда, стерео и/или синхронизированные многокамерные сетапы, которые позволяют рассматривать одно событие с разных точек зрения.
- Введение новых бенчмарков, сосредоточенных на понимании визуального опыта от первого лица в прошлом настоящем и будущем.



# Ego4D: Around the World in 3,000 Hours of Egocentric Video

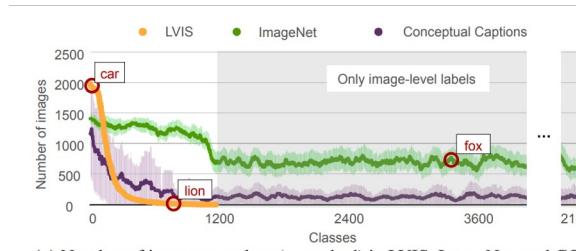


Демография людей, носящих камеры в наборе данных Ego4D.  
Включает возраст, пол, страны проживания и профессии.  
Размер шрифта отражает относительную частоту профессий.

Разнообразие датасета Ego4D в географическом распределении, видах деятельности и модальностях

# DETIC (Detecting Twenty-thousand Classes using Image-level Supervision)

- Обучение классификаторов детектора на данных классификации изображений
- Расширение словаря детекторов до десятков тысяч меток
- Отсутствие привязки меток изображений к bounding-box-ам на основе предсказаний модели
- Легкая реализация и совместимость с различными архитектурами детекции



(a) Number of images per class (smoothed) in LVIS, ImageNet, and CC.



(b) Results from an LVIS detector.

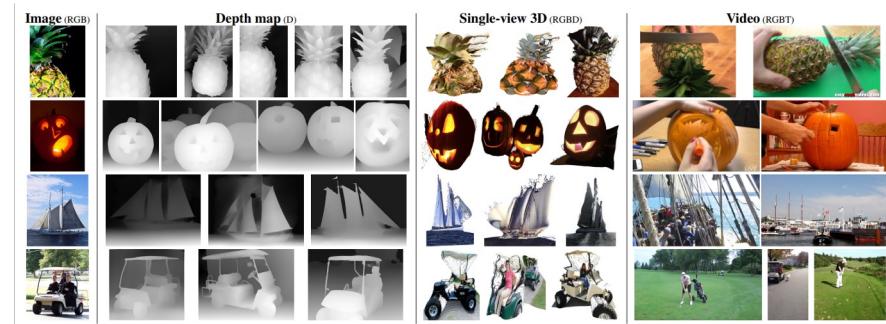


(c) Results from Detic.

Сравнение меток объектов у предыдущего детектора LVIS и DETIC

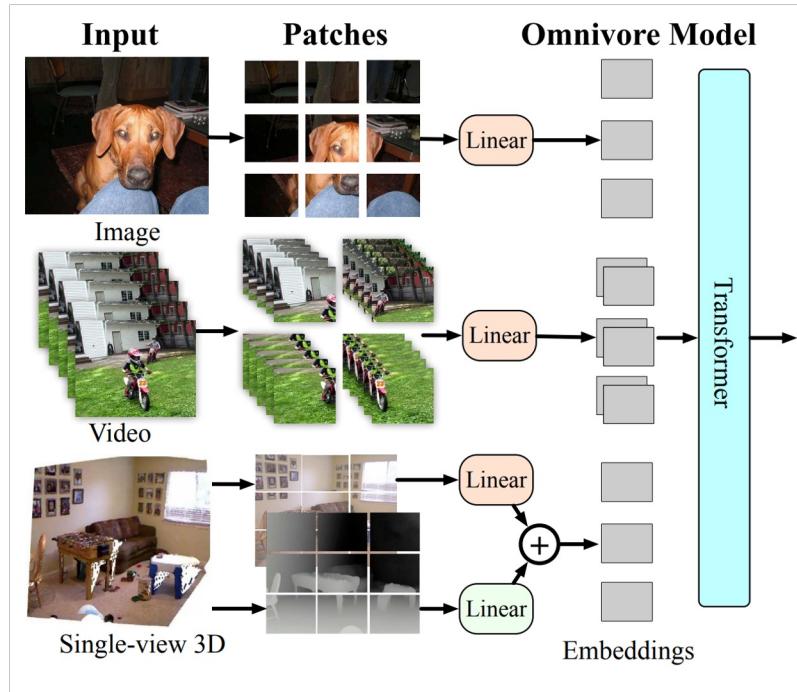
# Omnivore: A Single Model for Many Visual Modalities

- Omnivore способна классифицировать изображения, видео и данные 3D с одинаковыми параметрами модели
- Обучается совместно на задачах классификации из разных модальностей
- Преимущества
  - Прост в обучении
  - Использует стандартные наборы данных
  - Производит результаты на уровне или лучше моделей, специфичных для каждой модальности
  - Обобщается на различные модальности
- Кросс-модальное распознавание



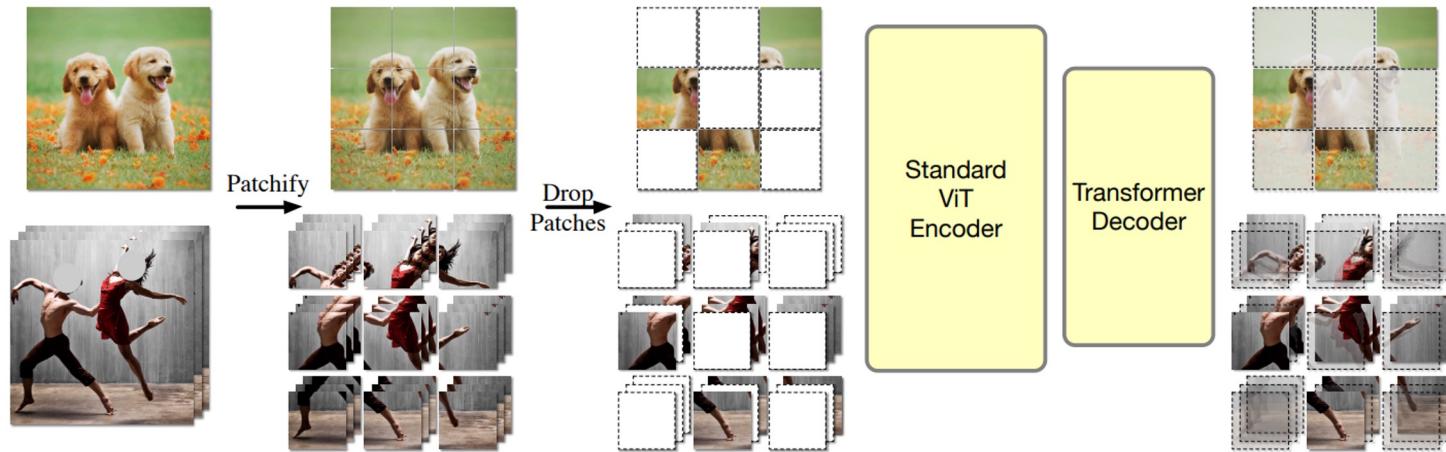
Данные, которые может принимать  
Omnivore

# Omnivore: A Single Model for Many Visual Modalities



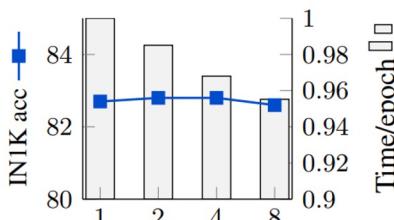
Архитектура Omnivore

# OmniMAE: Single Model Masked Pretraining on Images and Videos

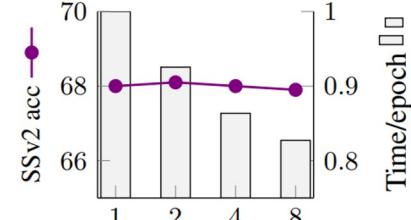


OmniMAE - единая модель для картинок и видео, которая обучается при помощи masked autoencoding

# OmniMAE: Single Model Masked Pretraining on Images and Videos

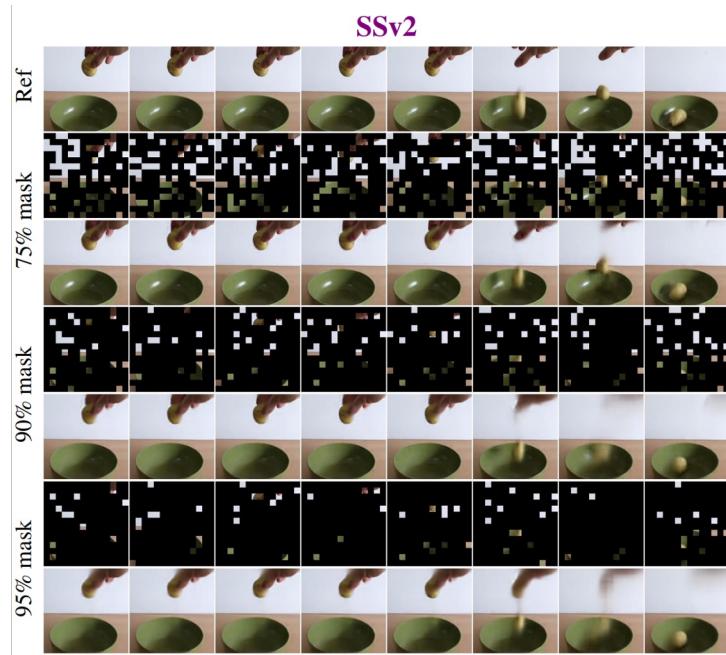


(a) Image sample replication



(b) Video sample replication

Если повторять в батче входные данные, то можно ускорить обучение модели без потери качества



Возможность восстанавливать видео при разном количестве замаскированных патчей

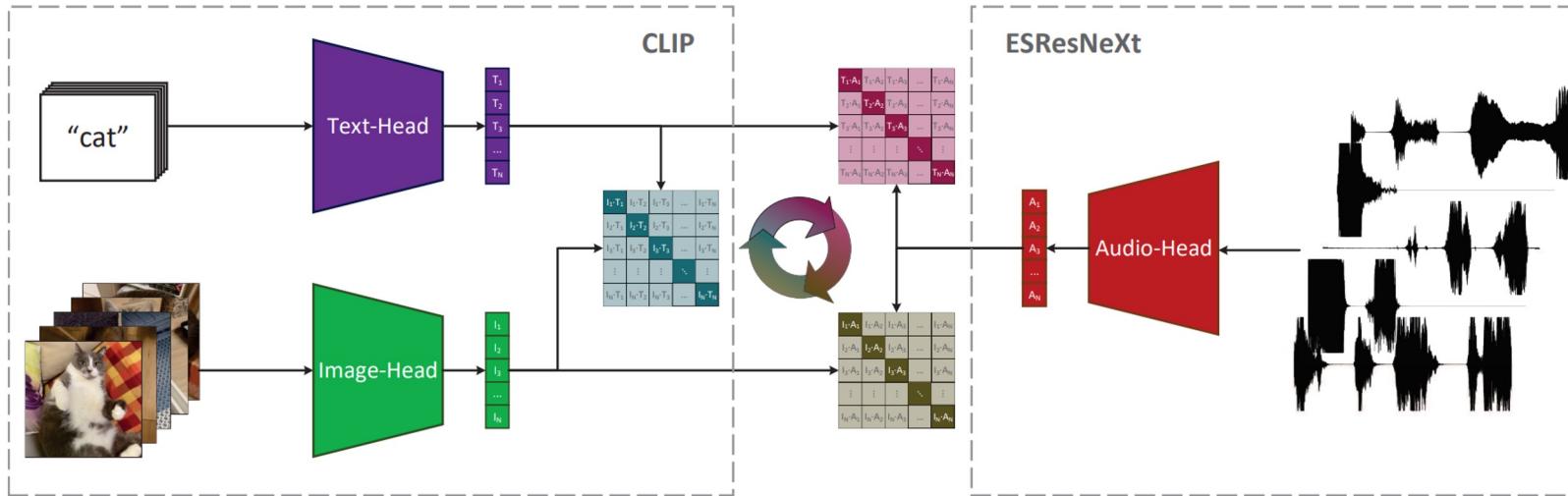
# Мультимодальные подходы до ImageBind

- AudioCLIP: Extending CLIP to Image, Text and Audio
- Learning Audio-Video Modalities from Image Captions
- Contrastive Multiview Coding
- Audio-Visual Instance Discrimination with Cross-Modal Agreement

# AudioCLIP: Extending CLIP to Image, Text and Audio

- AudioCLIP расширяет модель CLIP, добавляя обработку аудио к тексту и изображениям
- Предложенная модель объединяет аудио-модель ESResNeXt в фреймворк CLIP с использованием набора данных AudioSet
- Эта комбинация позволяет модели выполнять бимодальную и унимодальную классификацию и запросы, сохраняя способность обобщения на новые данные
- AudioCLIP достигает новых лучших результатов в задаче классификации окружающих звуков, превосходя другие подходы на UrbanSound8K и ESC-50.
- Модель устанавливает новый бейзлайн в задаче zero-shot прогноза ESC (Environmental Sound Classification)

# AudioCLIP: Extending CLIP to Image, Text and Audio

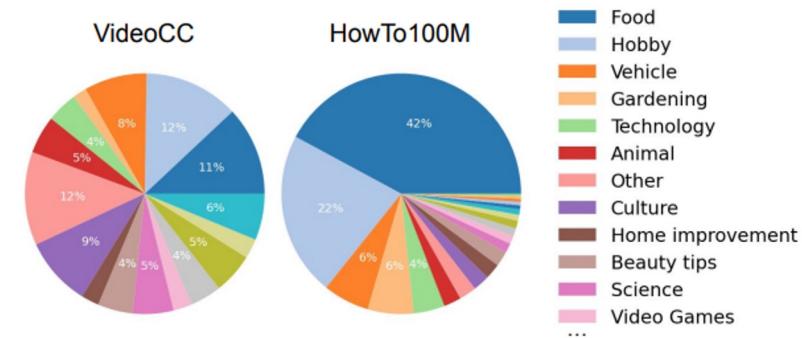


Архитектура AudioCLIP.

Слева показано обучение CLIP для текста и изображений. Справа показана аудио-модель ESResNeXT. Здесь добавленная звуковая модальность взаимодействует с двумя другими, позволяя модели обрабатывать три модальности одновременно.

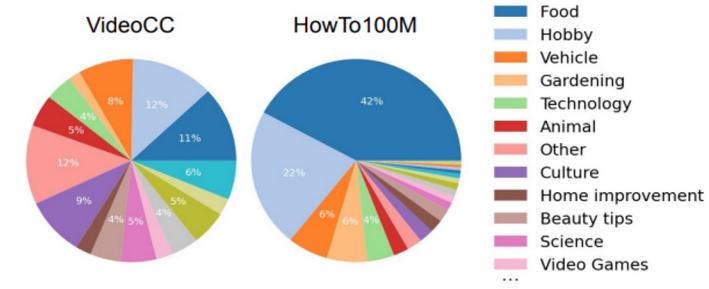
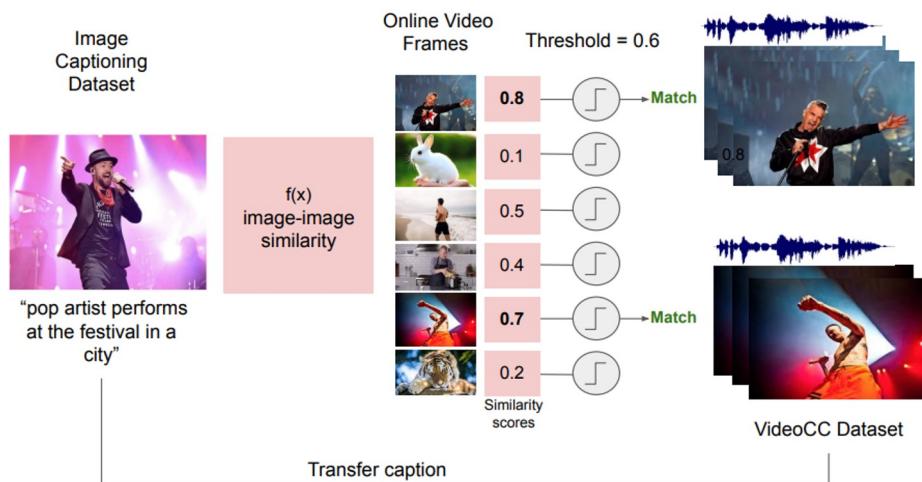
# Learning Audio-Video Modalities from Image Captions

- Одной из главных проблем в этих областях является отсутствие обширных обучающих данных.
- В статье предложен новый метод для сбора видео, в котором аннотации для видео и звука собираются на основе аннотаций картинок
- Модель, обученная на этом новом наборе данных, достигает конкурентоспособных результатов в поиске видео и генерации подписей для видео, сравнимых или превосходящих модель, предварительно обученную на наборе данных HowTo100M с значительно меньшим количеством видеороликов.



Сгенерированный в статье датасет намного более сбалансированный по сравнению с HowTo100M

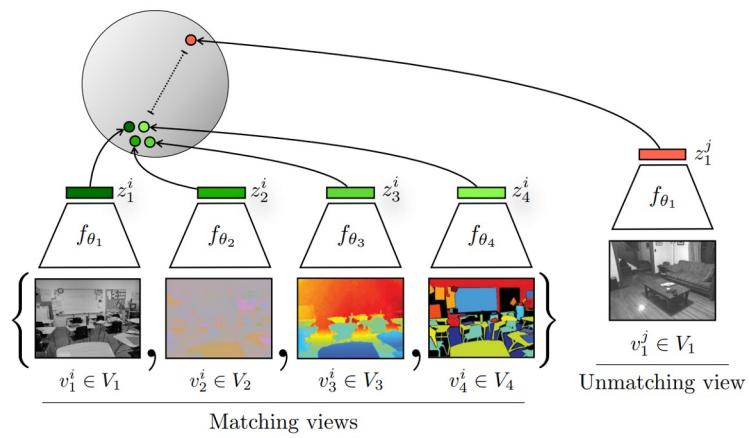
# Learning Audio-Video Modalities from Image Captions



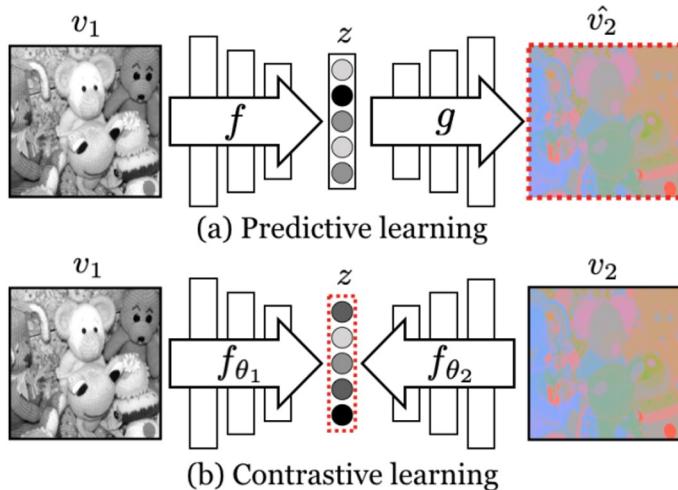
Сгенерированный в статье датасет намного более сбалансированный по сравнению с HowTo100M

Процесс присваивания меток для видео

# Contrastive Multiview Coding

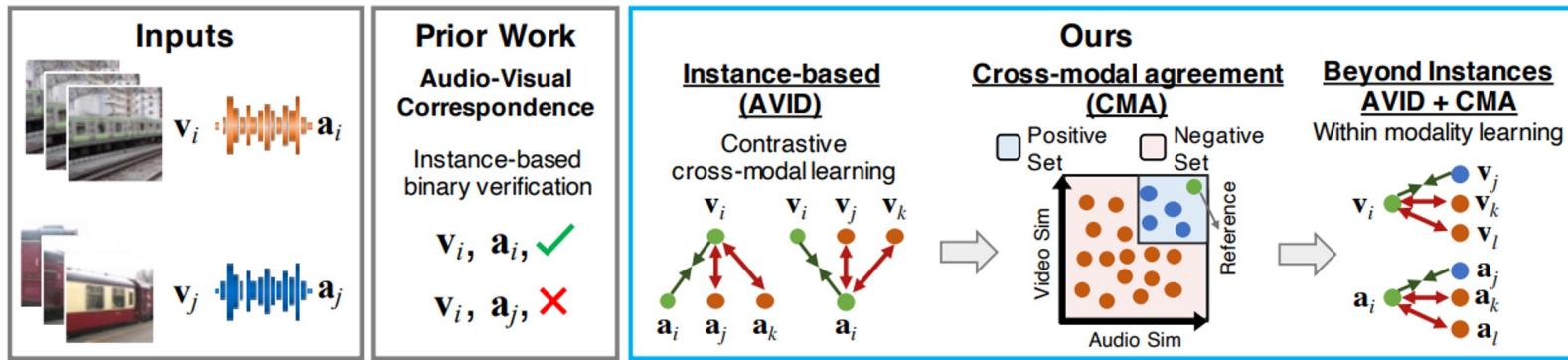


Применение contrastive learning для сеттинга с несколькими сенсорами. Хотим, чтобы репрезентации фреймов одного сеттинга были как можно ближе друг к другу, и дальше от фреймов остальных сеттингов



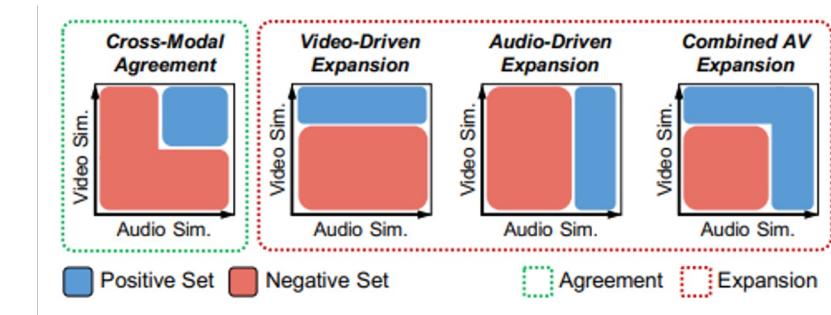
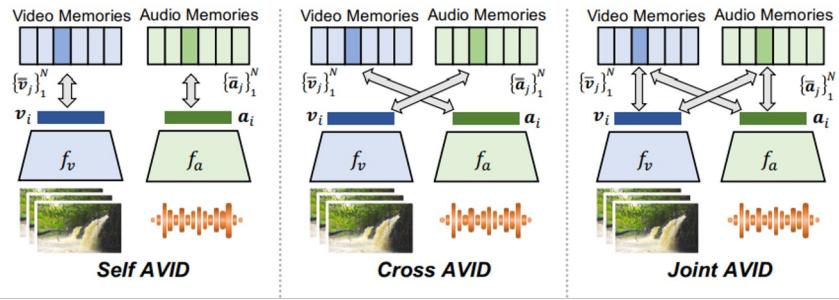
Авторы сравнивали два метода обучения:  
predictive learning и contrastive learning.  
Contrastive learning оказался более подходящим

# Audio-Visual Instance Discrimination with Cross-Modal Agreement



Авторы предложили последовательно применять AVID (Audio-Visual Instance Discrimination) и CMA (Cross-Modal Agreement) методы для мультимодального обучения

# Audio-Visual Instance Discrimination with Cross-Modal Agreement



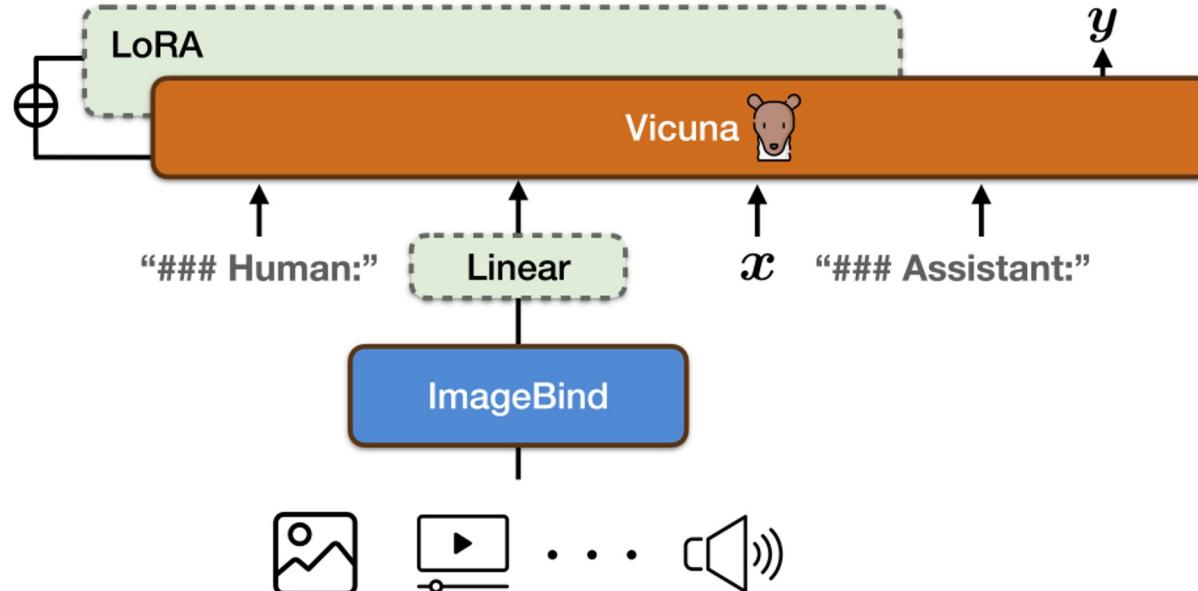
Были протестированы 3 похода к AVID.  
Лучшим оказался Cross-AVID

Обучения при помощи одного кросс-модального обучения.  
Также необходимо одномодальное. Для этого положительные  
примеры выбираются на основе схожести по одной или обеим  
модальностям. Лучшим оказался метод на основе схожести  
обеих модальностей.

# Модели использующие ImageBind

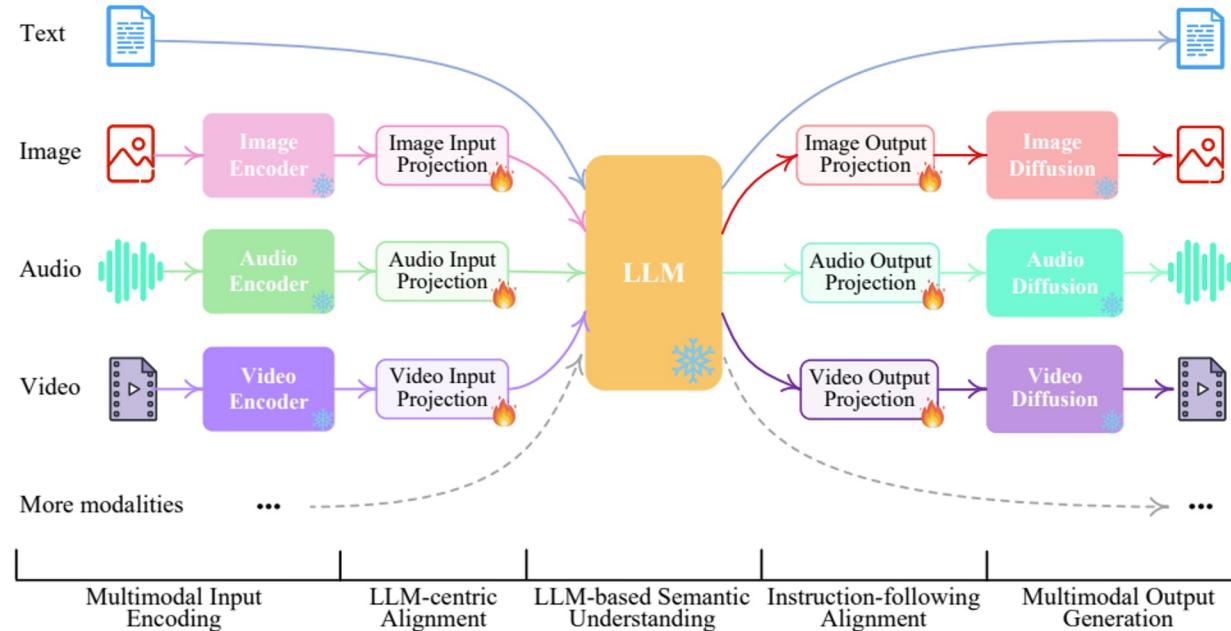
- PandaGPT: One Model To Instruction-Follow Them All
- NExT-GPT: Any-to-Any Multimodal LLM

# PandaGPT: One Model To Instruction-Follow Them All



Архитектура PandaGPT

# NExT-GPT: Any-to-Any Multimodal LLM



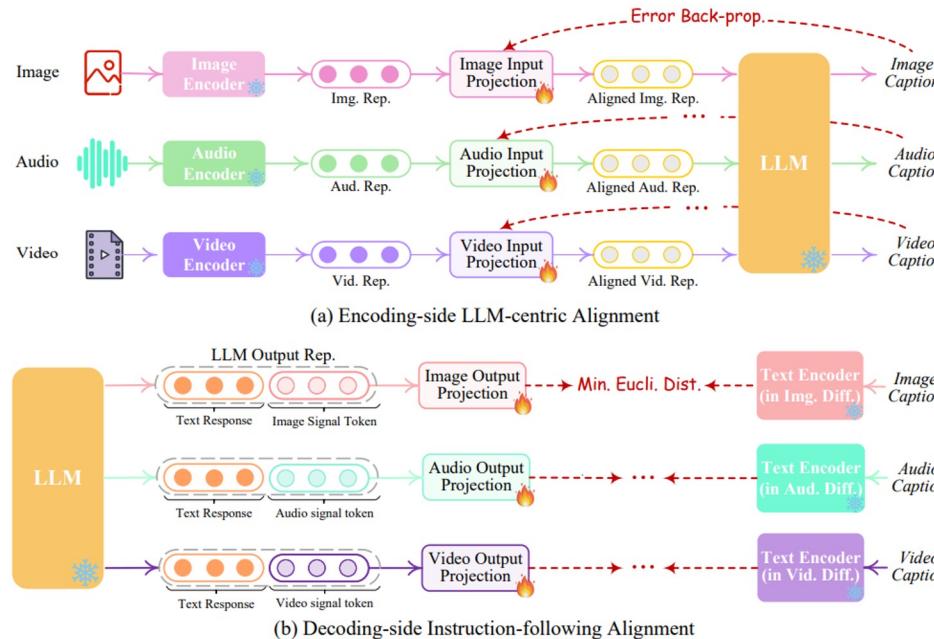
Из чего состоит NExT-GPT

# NExT-GPT: Any-to-Any Multimodal LLM

	Encoder		Input Projection		LLM		Output Projection		Diffusion	
	Name	Param	Name	Param	Name	Param	Name	Param	Name	Param
<b>Text</b>	—	—	—	—	Vicuna [12]	7B❄️	Transformer	31M🔥	SD [68]	1.3B❄️
<b>Image</b>	ImageBind [25]	1.2B❄️	Linear	4M🔥	(LoRA	33M🔥)	Transformer	31M🔥	AudioLDM [51]	975M❄️
<b>Video</b>							Transformer	32M🔥	Zeroscope [8]	1.8B❄️

Какие модели использовались и какие параметры  
обучались. Всего ~1% обучаемых параметров

# NExT-GPT: Any-to-Any Multimodal LLM



Процесс дообучения. Модули отвечающие за генерацию репрезентаций и мультимодальных данных дообучаются отдельно друг от друга