

Lyra: An Efficient and Speech-Centric Framework for Omni-Cognition

Supplementary Material

We strongly recommend that readers watch the video in our supplementary materials, which include more audio and video examples to get a better understanding and experience. In the following supplementary material, we provide more details about the training configurations and the construction and information of our dataset in Sec. A. In Sec. B, we present additional module settings along with some experimental results and analyses. In Sec. C, we showcase the qualitative results of Lyra.

A. Training Configuration and Data

A.1. Detailed Training Configuration

Stage-1: Speech Alignment. In this stage, we only train the parameters of the speech projector for speech-language pre-alignment with the LibriSpeech [48] and Common Voice Corpus [63] datasets, with about 1.0M data samples.

Stage-2: Joint Text-Image-Speech Training. Based on the Mini-Gemini [34] SFT data, we assemble and construct a unified dataset with 1.5M samples for the image-text-speech joint training. We use the ChatTTS [1] model to convert high-quality SFT data from text instructions into speech instructions. The multi-modal dataset, *i.e.*, Lyra-MultiModal-1.5M, includes not only single-turn instructions but also multi-turn instructions.

Stage-3: Long Speech SFT. To enable the model to integrate the long speech capability, we construct the first long-speech SFT dataset, called Lyra-LongSpeech-12K. Details can be found in Sec. 3.5 of the main paper. To ensure more robust performance, the dataset covers a wide range of topics, including humanities, social sciences, technology, education, and more. At this stage, we train both the speech module and the whole LLM module.

Stage-4: Streaming Text-Speech Generation. During the speech generation stage, we only train the speech generator. To better align the speech generator with the text decoder, we exclusively use text-speech modality QA pairs in our dataset. We filtered and selected a portion of suitable data from the datasets in our Stage-1, Stage-2, and Stage-3 for speech generation, resulting in a dataset of approximately 227K samples.

Detailed training settings are further explicated in Table 7.

Settings	Stage-1	Stage-2	Stage-3	Stage-4
Speech	Audio Length	< 30s	< 2500s, 30s clips	< 30s
	# Tokens	300	Max 25,000	300
Data	Dataset	LibriSpeech + CommonVoice	Lyra-MultiModal-1.5M	Lyra-LongSpeech-12K
	# Samples	1.2M	1.5M	12K
Training	Trainable	Projector	Projector + LLM	Projector + LLM
	Batch Size	256	128	16
	Learning rate	1×10^{-3}	2×10^{-4}	2×10^{-4}
	Epoch	1	1	3

Table 7. Detailed training settings of Lyra.

A.2. Data Collection and Curation

To ensure the data quality and training efficiency, we consider the following aspects while generating speech data for three modalities of joint training.

Generate multi-modal interleave data. To ensure models’ ability to process interleaved multi-modal data, we randomly select one round from multi-round conversations and convert its text into speech, while keeping the remaining rounds in text format. This guarantees that our SFT data preserves its multi-modal interleaved structure.

Oral Expression. Certain types of text are not well-suited for direct conversion using TTS technology. In these cases, we ensure the content is rewritten in a more conversational, oral form. For example, we rephrase “A:” as “Option A is” to enhance clarity and naturalness.

Speaker Diversity. To maintain diversity in our generated speech, we randomly select speakers with varying timbres and pitches for each instance. Since ChatTTS [1] obtains different speaker characteristics through various Gaussian sampling, it exhibits great diversity and robustness. During our generation process, we switch to a new set of ChatTTS random samples every 128 instructions.

Be Aware of the OCR Text. In real-world applications, a MLLM retrieves text by calling the OCR interface, such as TextVQA. Many OCR tokens, such as ‘G0’ and ‘EF’, lack clear meaning and are not suitable for verbal expression as speech input. Following this practice, we do not convert OCR text into speech.

Here, we list some training prompts and evaluation examples of our data in Fig. 7.

B. More Component-Wise Details & Analysis

B.1. Latent Multi-Modality Regularizer

Ablation of hyper-parameter λ in LCMR. In Table 9, we present the ablation study on the hyper-parameter λ . It can be observed that the performance is optimal when λ is set to 0.5, which consistently surpasses the results obtained without the application of LCMR, *i.e.*, $\lambda = 0$, across all speech-image benchmarks. Additionally, we note that setting this parameter too high can be detrimental to the overall performance.

Analysis Results on Speech/Audio Benchmarks. In Table 11, we have conducted evaluations on several renowned speech benchmarks, namely LibriSpeech [48] and AIR-Bench [72], where Lyra has achieved state-of-the-art results in speech capabilities. It surpasses all existing SLMs and omni models.

B.2. Latent Multi-Modality Extractor

Latent multi-modality extractor training performance. Qwen2-VL is exceptionally powerful, with the quantity and quality of its training data far surpassing those of public datasets and open-source models. As a result, most approaches to continual learning based on Qwen2-VL tend to result in performance degradation. Therefore, to evaluate the performance of our extractor module, we opt to train a new model from scratch. The results are shown in Table 8. Under the same training settings, models using latent multi-modality extractor achieve faster training speeds, with a maximum acceleration of nearly 50%. Additionally, they maintain or even improve average performance by up to 1% across multiple benchmarks. This series of experiments demonstrates the effectiveness of our extractor. Visualization of the latent multi-modality extractor in image modality is shown in Fig. 10. From it, the tokens retained in different blocks are all related to the user’s instruction. Additionally, for different questions, the token regions in the image most relevant to the question are preserved. This result is consistent with the video and speech modalities discussed in our main paper.

Compared with FastV [8]. LMME can progressively discard unimportant context tokens across multiple layers dynamically, whereas FastV is quite rigid, discarding them all at once after a constant layer. Moreover, LMME applies to a wider range of modalities, such as audio and long-speech cases, not just the vision modality. The detailed comparison is shown in Table 10.

Prefill time, TPS, memory, TFLOPs comparison. In Table 12, we vary the token length, ranging from 2^{11} to 2^{17} (under a long-context case). We denote LMME(n , ρ) as splitting the LLM into n blocks, with each block retaining the top ρ proportion of the most important tokens. We compare three models: the baseline, LMME(4, 0.8), and LMME(4, 0.7). The key metrics examined include Prefill Time, tokens-per-second (TPS), and memory usage on the A100 GPUs. Under the baseline model, multimodal content exceeding 2^{15} tokens results in out-of-memory (OOM) errors. In contrast, our models LMME(4, 0.8) and LMME(4, 0.7) still have room for 2^{17} tokens, consuming over 50% less memory. Additionally, the Prefill Time is significantly shorter than the baseline model (by 100%), and the token generation speed is also notably faster (by 50%). Additionally, FLOPs reduce about 50% in most cases.

LMME training time comparison on multi-modality datasets. In Table 14, we primarily examine the improvement in training speed. We evaluate it using our proposed Lyra SFT and long-speech SFT dataset, which contains 1.5M samples and 12K samples, respectively. From the table, our LMME can reduce training time by more than 50% compared to the original. Since the context in the long-speech dataset is generally longer than it is in the 1.5M dataset, the acceleration effect becomes even more pronounced.

Method	LLM	Vision	Data	Time	TextVQA	MME	MM-Vet	MMB-EN	SEED	MMMU	Avg. Rate
Baseline	Vicuna-7B	CLIP+Conv	Lyra-MM-1.5M	65h	68.4	1865	41.3	65.8	68.1	36.8	100.0%
+ Extractor	Vicuna-7B	CLIP+Conv	Lyra-MM-1.5M	35h(-46%)	69.9	1899	44.9	66.7	67.5	35.3	101.5%(+1.5%)
Baseline	Qwen2-7B	SigLIP	LLaVA-665K	18h	69.7	1974	39.4	76.7	74.2	40.8	100.0%
+ Extractor	Qwen2-7B	SigLIP	LLaVA-665K	14h(-22%)	69.1	2005	38.6	76.9	73.5	40.6	99.6% (-0.4%)
Baseline	Qwen2-7B	SigLIP	Lyra-MM-1.5M	51h	71.9	2030	51.0	78.1	74.5	40.2	100.0%
+ Extractor	Qwen2-7B	SigLIP	Lyra-MM-1.5M	35h(-31%)	71.8	2007	50.6	77.7	73.7	42.1	100.1%(+0.1%)

Table 8. **Latent multi-modality extractor training performance.** The training time is reduced by an average of **one-third**, while the average performance does not degrade and even **improves by 0.4%**.

Hyper-parameter λ	TextVQA ^S	MM-Vet ^S	ChartQA ^S	AI2D ^S
Baseline ($\lambda = 0.0$)	79.0%	59.5%	58.8%	64.7%
Lyra (LCMR, $\lambda = 0.3$)	79.8%	60.1%	59.4%	65.8%
Lyra (LCMR, $\lambda = 0.5$)	80.0%(+1.0%)	60.5%(+1.0%)	60.4%(+1.6%)	66.4%(+1.7%)
Lyra (LCMR, $\lambda = 1.0$)	78.8%	58.9%	57.2%	63.8%

Table 9. **Ablation of hyper-parameter λ in LCMR.** Bench^S indicates that it uses speech instruction as the input.

Model	ASR on LibriSpeech [↓]		AIR-Bench [↑] [72]
	test-c	test-o	Chat-speech
SpeechGPT [78]	-	-	1.57
Whisper-small [51]	4.4	10.1	-
SALMONN [56]	2.1	4.9	6.16
Qwen2-Audio [11]	1.6	3.6	7.18
LLaMA-Omni [16]	-	-	5.22
Mini-Omni2 [69]	4.8	9.8	3.58
VITA-1.5 [19]	3.3	7.2	4.83
IXC2.5-OmniLive [79]	2.5	5.7	1.60
Lyra-Base	1.8	3.8	7.51

Table 11. **Analysis Results on Speech/Audio Benchmarks.**

Data Effectiveness	TextVQA ^S	DocVQA ^S	ChartQA ^S	AI2D ^S
InterOmni (27M data)	69.0%	80.0%	56.1%	54.0%
Lyra (Lyra-HD-0.7M)	79.0%	82.0%	56.5%	64.1%
Lyra (Lyra-HD-1.5M)	80.0%(+1.0%)	84.6%(+2.6%)	60.4%(+3.9%)	66.4%(+2.3%)

Table 13. **Validation of the proposed dataset effectiveness.**

Methods Comparison	FLOPS	TextVQA	MMBench	MME
Baseline	100%	82.6	79.0	2326
FastV ($K=2, R=0.3$)	55%	80.1	76.3	2155
LMME ($n=4, \rho=0.6$)	55%	80.2(+0.1%)	77.8(+1.6%)	2283(+128)
FastV ($K=2, R=0.3$)	70%	81.4	78.0	2280
LMME ($n=4, \rho=0.8$)	65%	82.0(+0.6%)	78.8(+0.8%)	2286(+6)

Table 10. **Comparison of FLOPs and performances with FastV.**

Metric	# (Tokens)	2 ¹¹	2 ¹²	2 ¹³	2 ¹⁴	2 ¹⁵	2 ¹⁶	2 ¹⁷
Prefill(s) [↓]	Baseline	0.19	0.33	0.65	1.47	2.99	OOM	OOM
	LMME(4, 0.8)	0.17	0.24	0.44	0.76	1.60	4.24	10.2
	LMME(4, 0.7)	0.16	0.21	0.37	0.59	1.23	3.05	7.75
TPS [↑]	Baseline	32.6	30.8	27.3	25.3	16.6	OOM	OOM
	LMME(4, 0.8)	32.7	31.5	31.8	28.6	22.7	14.1	8.37
	LMME(4, 0.7)	33.8	33.3	32.5	30.1	25.3	16.6	10.1
Memory [↓]	Baseline	20G	23G	30G	41G	60G	OOM	OOM
	LMME(4, 0.8)	17G	18G	19G	21G	24G	33G	49G
	LMME(4, 0.7)	17G	18G	19G	21G	24G	33G	49G
TFLOPs [↓]	Baseline	17.5	37.3	74.3	156	287	OOM	OOM
	LMME(4, 0.8)	8.63	19.1	47.6	96.2	183	393	785
	LMME(4, 0.7)	7.42	16.4	33.8	84.4	157	337	674

Table 12. **Prefill time, TPS, memory, TFLOPs comparison.**

Data Type	Baseline	LMME(4, 0.9)	LMME(4, 0.8)	LMME(4, 0.7)
Lyra-MM-1.5M	66h	58h (-13%)	47h (-29%)	41h (-38%)
Lyra-LongSpeech-12K	9.6h	7.0h (-27%)	5.7h (-40%)	4.5h (-54%)

Table 14. **LMME training time comparison.**

B.3. Long Speech Capability Integration

Additionally, we introduce prompts related to the long speech capability. The detailed prompts are shown in Fig. 8. The first is the GPT-4o-based prompt used to generate Q&A during the long speech data collection process. The second is the inference prompt we used to apply the long-speech Lyra model on the VideoMME benchmark. For detailed results and analysis, refer to Sec. 3.5 and the long-speech capability integration part in Sec. 4.3.

B.4. Sound Capability Integration

For the sound modality, due to the lack of many pretrained models, we primarily follow ImageBind[21] as the sound encoder. ImageBind processes sound, text, and image modalities using a training approach similar to CLIP [50], ultimately encoding them into just one single token. This approach is not particularly generalizable. During the sound SFT process, our model based on LLaMA3 [15] is trained on the AudioCaps [29] dataset, which contains a total of 46K training samples. The quantitative performance of our model on the test set is shown in Table 15.

Lyra Data Examples

Training conversations:

human: <image>\nWhat are the two people holding?\nAnswer the question using a single word or phrase.

GPT: Umbrella.

human: What is the person with the Red Hat doing?

GPT: Taking pictures.

human: <speech>

GPT: Blanket.

Evaluation cases:

human: <image>\nReference OCR token: DAKOTA, DIGITAL, Single-Use, Camera, Pire, digitat\n<speech>

Figure 7. **Lyra training and evaluation data examples.**

Long Speech Question-Answer Generation Prompt Example

Task:

You will be provided with a transcript from an audio or video recording. Your task is to generate question-answer pairs based on the content of the transcript.

Guidelines for Question-Answer Pair Generation:

- The first question should be about summarizing the content of this recording.
- Carefully read the transcript provided and base all questions and answers strictly on the content within.
- Ensure that each question is directly related to specific details in the transcript, such as events, facts, or points made by the speaker.
- Provide clear, concise, and specific questions, along with accurate answers derived from the transcript.
- Do not introduce any new information that isn't in the transcript. If the speaker does not introduce themselves, refer to them as "Speaker" or "Narrator".
- Avoid generic or overly broad questions; aim for a range of question types (e.g., factual, inferential, explanation-based).
- Generate five question-answer pairs.

Output Format:

- Your output should be structured as a JSON object.
- Each question-answer pair should be formatted as:

```
```\njson\n{\n  [\n    {\n      "Question": <question-1>, "Answer": <answer-1>,\n      "Question": <question-2>, "Answer": <answer-2>,\n      ...\n    }\n  ]\n}\n```\n
```

### Long Speech VideoMME Evaluation Prompt Example

Based on the context, determine if it provides enough information to answer the question:

<question> with the provided choices <option-A>, <option-B>, <option-C>, <option-D>.

Do not introduce any information not found in the context.

- If the context is sufficient to answer the question, respond "yes" and answer with the option's letter from the given choices directly.
- If the context does not contain enough information to answer the question, respond "no".

Figure 8. **Long speech related prompt examples.**

Regarding this dataset, as the authors of AudioCaps [29] have noted, "Even to humans, recognizing the true identity of a sound can be ambiguous." Moreover, LLM-based multimodal models tend to produce more detailed descriptions, while metrics like SPICE [3] and CIDEr [62] are outdated and fail to effectively reflect the most suitable results. Even under such circumstances, our Lyra, trained on just 46K samples for the sound modality, outperforms previous sound models. Some qualitative results are shown in Fig. 9.



Figure 9. Sound capability qualitative results.

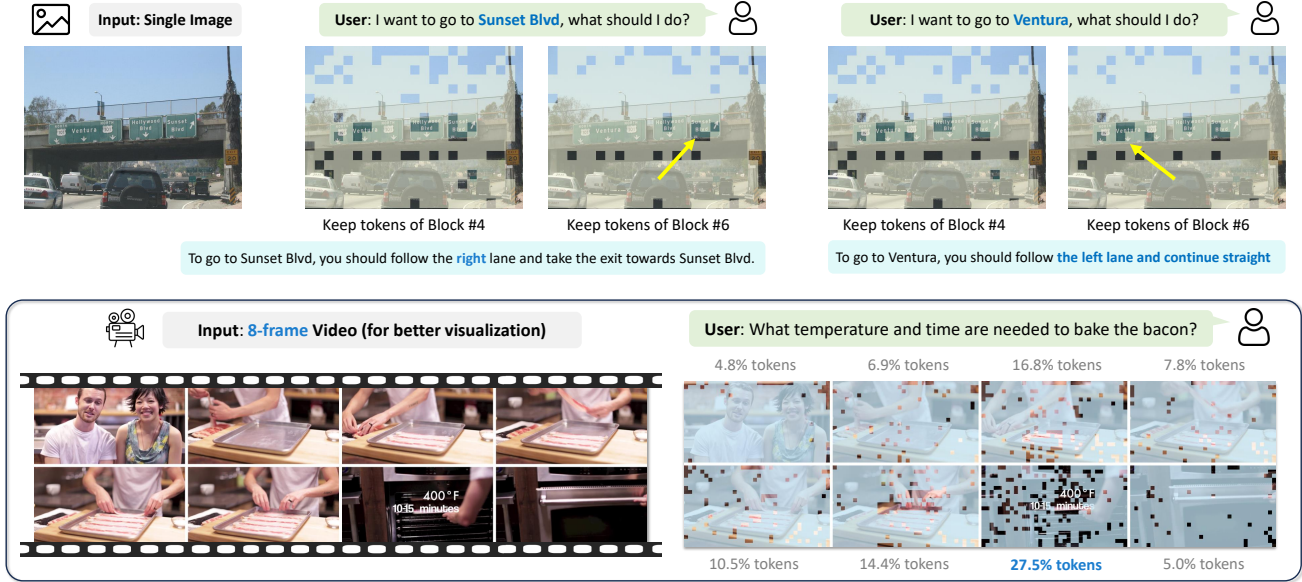


Figure 10. Visualization of latent multi-modality extractor in the image (upper) and video (bottom) modality.

## B.5. Streaming Text-Speech Generation

For the speech-text streaming speech generation component, **we have adopted two approaches: non-autoregressive (NAR) and autoregressive (AR)**. The non-autoregressive mode is inspired by LLaMA-Omni [16], offering lower latency (A lag of about 0.5 seconds.), but with a slight compromise in sound quality. The autoregressive mode, on the other hand, is based on Mini-Omni [84], providing better control over the characteristics of the generated speech. **We support both English and Chinese speech output in this mode.**

**Speech Discretization.** To handle NAR speech responses, we discretize the audio into discrete units with the following steps: 1). Continuous representations are extracted using the HuBERT model [25]. 2). These representations are clustered into discrete indices via the K-means algorithm. 3). Consecutive repeated indices are merged to form a sequence of discrete units, which can be converted back to waveforms using a vocoder [49]. To handle AR speech responses, we discretize the audio into discrete units by lightweight SNAC [55] encoder. It uses the downsample factors (or strides) of [8, 4, 2, 1]. Each codebook holds 4096 entries (12-bit).

**Speech Decoder for Streaming Generation.** A streaming speech decoder is introduced after the LLM to enable simultaneous generation of text and speech: For the NAR mode, to ensure the overall structure remains consistent with the LLM, the decoder is built using **two transformer layers** similar to Qwen2-VL [64]. Similar to LLaMA-Omni, it processes the hidden states from the LLM and generates discrete speech units in a NAR manner [41, 81]. For upsampling, the text hidden states from the LLM are upsampled to match the speech sequence’s length. These upsampled representations are processed by the speech decoder to produce output features for the discrete speech units. Due to the increased complexity of encoding in AR mode, we employ **4 to 6 transformer layers** to process the AR encoding.

**Alignment and Training.** For the NAR mode, following LLaMA-Omni, Connectionist Temporal Classification (CTC) [24] is used to align the decoder’s output with the discrete speech units. During training, the model learns to match the output features to the target speech units by minimizing the CTC loss. During inference, the most likely sequence is selected, converted into discrete units, and passed through the vocoder to generate audio. For the audio and text tokens generated simultaneously, the negative log-likelihood loss is adapted in the AR mode training process.

AT [43]	BART [22]	PairMix [30]	CoDi [58]	Lyra-Base
16.8	17.7	18.1	17.1	19.5

Table 15. Sound SPICE performance comparison.

Eval/Train	ChatTTS	Edge-TTS	Eval/Train	ChatTTS	Eval/Train	ChatTTS
ChatTTS	<b>80.0</b>	<b>79.5</b>	ChatTTS	84.6	ChatTTS	60.4
Edge-TTS	79.7	78.3	Intern-O	82.3	Intern-O	58.3

(a) TextVQA<sup>S</sup>

(b) DocVQA<sup>S</sup>

(c) ChartQA<sup>S</sup>

Table 16. Different TTS training and evaluation.

Listing 1. Sample Random Function in ChatTTS (Pytorch)

```

1 def sample_random(self) -> torch.Tensor:
2 spk = (
3 torch.randn(self.dim, device=self.std.device, dtype=self.std.dtype)
4 .mul_(self.std)
5 .add_(self.mean)
6)
7 return spk

```

## B.6. TTS Methods Ablation Study

In this subsection, we briefly compare the impact of different TTS (text-to-speech) methods on the generalization and robustness of speech instruction (across different domains). We primarily used two TTS methods: ChatTTS [1] and Edge-TTS [44]. ChatTTS employs Gaussian sampling to simulate different speakers (As shown in Listing 1), while Edge-TTS randomly selects from a fixed set of 41 speakers. ChatTTS is likely to be more diverse. We trained models using instruction data generated by these TTS methods and evaluated TextVQA speech instructions generated by different TTS methods. Detailed results can be found in Table 15a. Models trained with speech generated by ChatTTS demonstrated better generalization due to its diversity.

Similar results were observed when compared with speech instructions generated by Intern-Omni [47]. Because we cannot access their training speech instruction data; they only provided the evaluation speech instruction data of DocVQA and ChartQA. Specific results are provided in Table 15b and 15c. While models perform better when trained and evaluated on instructions generated by the same system, the experiments overall demonstrate that instructions generated by ChatTTS are more robust compared to the other two methods.

## C. Qualitative Results

### C.1. Examples of Images and Videos


In Fig. 11, we present additional interactions with Lyra, showcasing the model’s adeptness in knowledge-based perception and reasoning for both images and videos. In various complex scenarios, such as recognition of complex PC backgrounds, understanding of game interfaces, and analyzing football match videos with significant differences between frames, Lyra demonstrates superior understanding and reasoning cognitive outcomes.

### C.2. Examples of Long Speeches

In the main paper experimental section, Fig. 6a shows that existing Speech Language Models (SLMs) fail when processing audio longer than 450 seconds (about seven minutes): the output becomes nonsensical with extensive repetition. In this part, we demonstrate Lyra’s ability to handle long audio inputs. In Fig. 12, 13, 14, and 15 we demonstrate Lyra’s capability to process long-form speech ([best view the following part together with the video in our supplementary materials](#)). Lyra effectively extracts the information that users need from extended speech contents. It excels at capturing both the details and the overall structure of long speeches. In news scenarios (Fig. 12, with frequent topic switches), it accurately identifies the focused information and responds exceptionally well.


For more complex tasks, as shown in Fig. 13, such as scenarios with visual ambiguity, our model leverages long-form speech and keyframes from videos to provide more accurate results compared to a powerful VLM like Qwen2-VL that rely solely on visual information. In Fig. 14, our model demonstrates its ability to process daily lectures, offering significant advantages for educational-related applications. Lyra can handle speech content durations exceeding two hours, which enables intelligent models to tackle more complex multi-modal tasks. In Fig. 15, For tasks with longer temporal sequences and higher complexity, Lyra can also understand them and provide subjectively reasonable answers to the questions.

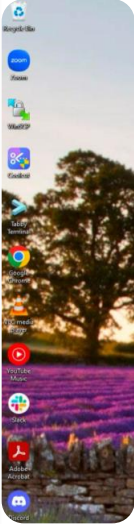




**Input: Single Image**

How many APP icons are in this image?  
List their names and the corresponding functions in two columns.





There are 10 APP icons visible

Name	Function
Recycle Bin	A container for temporary....
Zoom	A video conferencing ...
WinSCP	A secure file transfer ...
CoolTerm	A terminal emulator...
Tabby Terminal	A terminal emulator ...
Google Chrome	A web browser for accessing ...
VLC media player	A multimedia player ...
YouTube Music	A music streaming
Slack	A collaboration platform ...
Adobe Acrobat	A PDF document viewer ...

Lyra-Base-9B

There are 12 APP icons

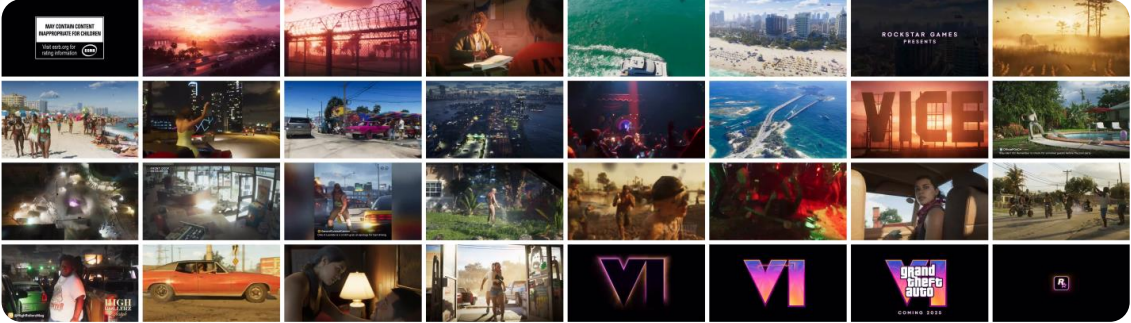
Name	Function
Recycle Bin	A place to store files ....
Zoom	A tool for magnifying the screen
Windows	The operating system...
Google Chrome	A web browser for browsing ...
Google	A search engine for finding...
Adobe Acrobat	A program for viewing ...
YouTube	A video-sharing platform...
Slack	An app designed for teams ...
Discord	A voice and text chat app ...
Tabs	A browser extension...
ARC media player	A media player for playing ...
Google Drive	A cloud storage service...

Mini-Gemini-HD-34B

11 APP icons


Name
Recycle Bin
Zoom
WinSCP
Coolcut
Tabby Terminal
Google Chrome
VLC media player
YouTube Music
Slack
Adobe Acrobat
Discord

GT




**Input: 2 min Video**

User: Please describe this video.




The video showcases a series of scenes from a video game, including a woman in a bikini walking on the beach, a man in a car crashing into a store, and a woman being chased by a group of men on motorcycles. The video also features a logo for the game "Grand Theft Auto VI" and a warning that the game may contain content inappropriate for children. The scenes are set in a city with palm trees and a beach, and the game appears to be a crime-themed action game.



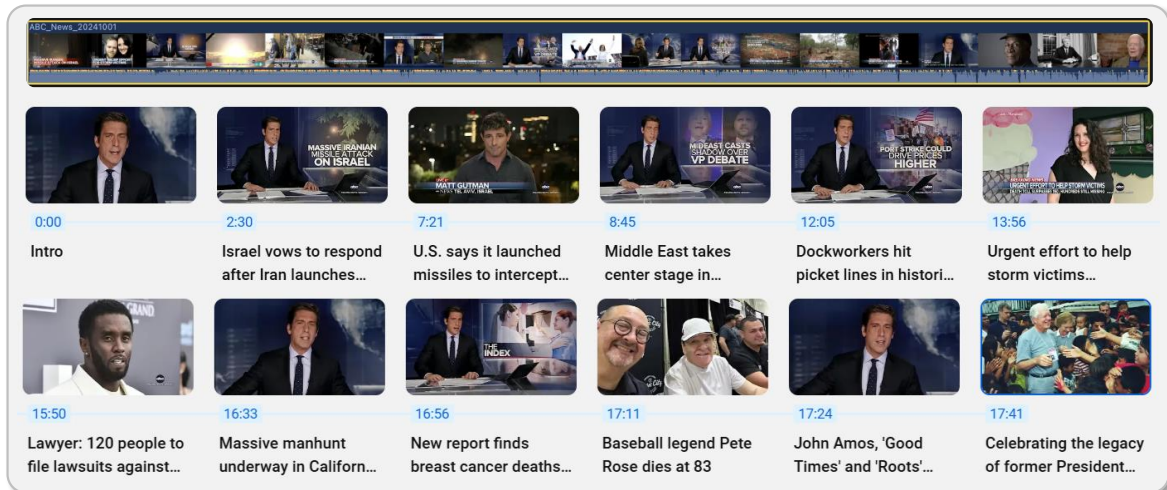
**Input: 15 sec Video**

User: Provide a detail commentary for this football match clip.



The soccer match is between **England and Iran**, and the game is currently at the **34th minute**. The teams are playing on a field with a large crowd watching. The players are actively engaged in the game, with **one player from England scoring a goal**. The goalkeeper from Iran is unable to catch the ball, and the crowd cheers for the goal. The match is being broadcasted by FIFA.

Figure 11. Image-text and video-text qualitative results of Lyra.



**Input: 20min 1s Audio**

**User:** Give me a summary about today's news.

The recording covers multiple significant events and news stories.

- It begins with **Iran launching over 200 ballistic missiles** at Israel in response to the killing of Hezbollah leader in Beirut, causing air raid sirens and explosions across Israel, with Israeli defenses and US military intercepting many missiles.
- The situation raises fears of a wider regional war, affecting the US presidential race with the **vice-presidential debate** looming.
- Meanwhile, in the US, **tens of thousands of dockworkers are on strike**, threatening the national supply chain, while search efforts for victims of **Hurricane Helene continue with over 150 deaths** reported.
- There are also developments in legal cases against **attorney Shaun Diddie Combs and a manhunt for an escapee**.
- The recording also covers the **passing of baseball great Pete Rose and actor John Amos**, along with updates on **President Jimmy Carter's 100th birthday**.

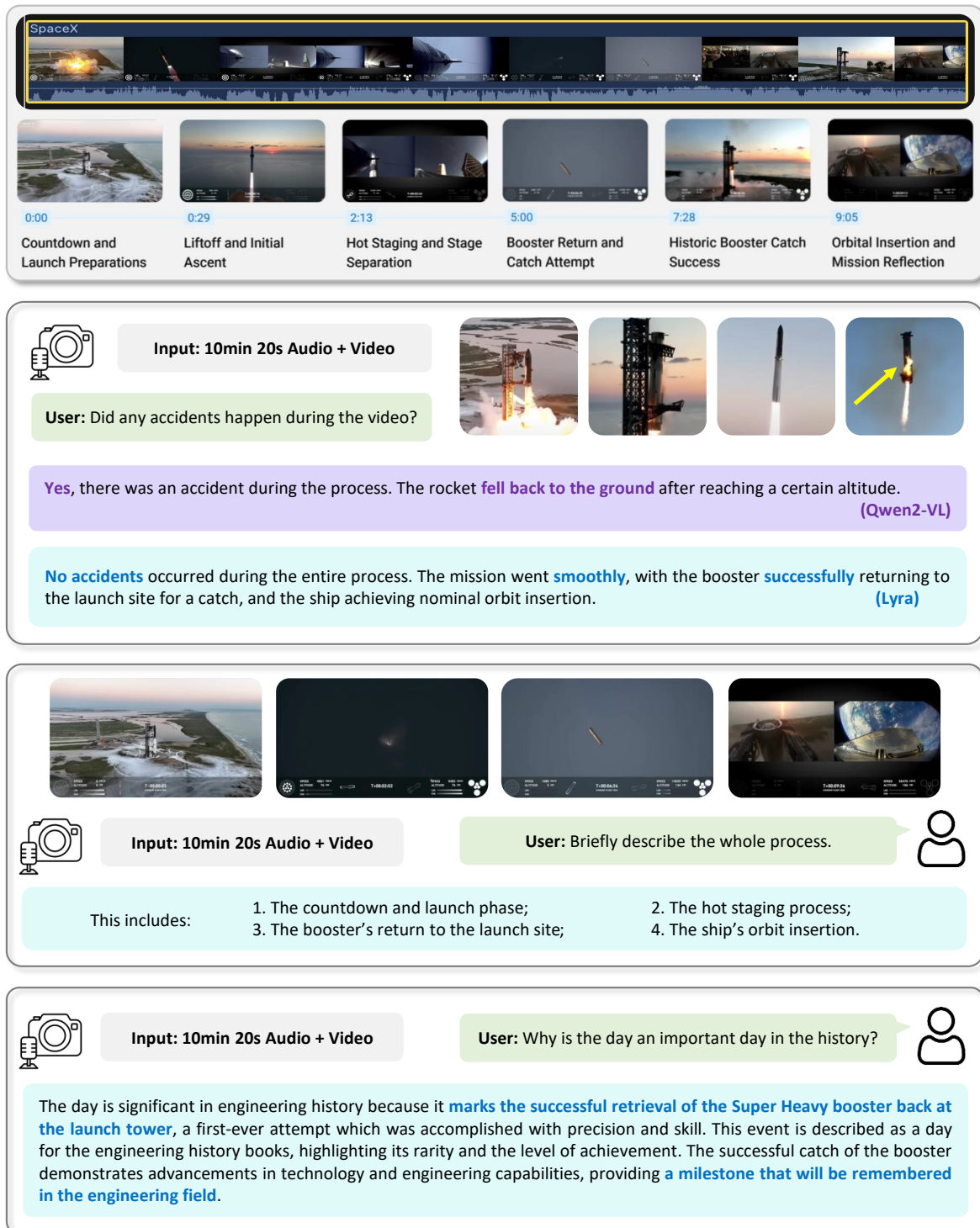
**Input: 20min 1s Audio**

**User:** Give me more information about the Hurricane Helene.

Hurricane Helene was a significant hurricane that struck the northeastern United States, particularly affecting areas such as New England and the Mid-Atlantic states. The hurricane made landfall on the coast of Maine and brought heavy rainfall, strong winds, and storm surges that caused extensive damage. The death toll from the hurricane rose to more than 150 lives, and search and rescue operations were ongoing as teams worked through the night to locate missing individuals. The disaster also caused significant destruction, with rescuers searching from structure to structure, including finding a couple who had been swept away by a wall of water. The search efforts were described as growing more desperate by the hour, with a focus on locating one man's fiancée who was still missing.

Figure 12. Lyra long speech capability qualitative results for handling daily news.





0:00  
Introduction example

1:07  
Series preview

2:42  
What are neurons?

3:35  
Introducing layers

5:31  
Why layers?

8:38  
Edge detection example

11:34  
Counting weights and biases

12:30  
How learning relates

13:26  
Notation and linear algebra

15:17  
Recap

16:27  
Some final words

17:03  
ReLU vs Sigmoid



Input: 18min 39s Audio

User: Give me more other activation functions **not mentioned in today's lecture**, and give me **the mathematical expression**.



According to the video lecture, neural networks have shifted from using **sigmoid** functions, which compress inputs to values between 0 and 1, to the simpler **ReLU** function, which outputs either 0 or the input value itself, offering more efficient neuron activation.

Other activation functions that have been used or proposed include **Leaky ReLU**, which allows a small, non-zero gradient when the neuron is not activated, and **ELU** (Exponential Linear Unit), which uses an exponential function to provide negative values for inputs below zero. These functions help introduce non-linearity into the network, enabling it to learn more complex patterns.

**Leaky ReLU:**  $\max(0, x) + \alpha \min(0, x)$

4:50  
Origin of stars

9:41  
Planetary nebulae

12:29  
Interstellar gas and its properties

16:25  
Studying interstellar gas

20:21  
Star formation and the interstellar medium

30:09  
Theory of star formation

35:53  
Birth of stars

40:57  
Observing star formation

42:56  
Formation of planets

48:07  
Star formation

51:53  
Evaporation of star clusters

1:11:57  
Radioactivity and the nuclear reactions

1:23:57  
Neutrinos and their role in the life of stars

1:36:52  
Classification of stars

1:36:46  
Evolution of the Sun

1:43:44  
Pulsating stars

1:46:31  
Final stages of a star's life

1:48:24  
White dwarfs

1:53:57  
Supernova explosions

2:02:16  
Neutron stars and black holes

2:04:18  
Q&A session. Fate of living beings and...

2:04:59  
Planets colonization



Input: 2h 27min 55s Audio

User: What is the topic of today's discussion?



- Today's discussion is about the **evolution and formation of stars within the universe**.
- Initially, the speaker discusses the shift in understanding from stars being the primary focus of astronomy to recognizing the significant roles played by **dark energy and dark matter**.
- The discussion covers the **origins of stars, the processes involved in their formation, and the various environments in which they are born**.
- The speaker explains how cold molecular clouds and the interstellar medium play crucial roles in condensing into stars.
- The talk also delves into the **lifecycle of stars**, including their eventual transformation into planetary nebulae and the role of magnetic fields in their evolution.
- Furthermore, the speaker highlights the challenges and advancements in observing and studying these phenomena, including the use of radio and millimeter telescopes.
- The discussion concludes with insights into the formation of star clusters and the dynamics of stars moving through the galaxy.

Figure 14. More examples of Lyra with hour-long lectures (more than two hours).



Figure 15. More results from long speech examples: Lyra can subjectively answer questions about complex steps.