Appendix

A. More on Speech Tokenizer

A.1. Speech-to-Unit (S2U) Tokenizer

Overview. To process the speech input $\mathbf{X}_S$, our S2U tokenizer consists of a speech encoder $s(\cdot)$ with a quantization module $q(\cdot)$. First, the speech input is passed through $s(\cdot)$, producing a continuous latent representation $\mathbf{E}_S = s(\mathbf{X}_S)$. Then, the quantization module $q(\cdot)$ converts $\mathbf{E}_s$ into discrete units $\mathbf{U}_S = q(\mathbf{E}_s)$. The final output is an ID sequence $\mathbf{U}_S = [u_1, u_2, \cdots]$, where each u_i corresponds to a unique speech unit in the speech codebook.

After this S2U extraction procedure, the speech is represented by quantized IDs instead of quantized embeddings. For example, a speech signal is represented as an ID sequence like $[782, 463, 550, \cdots]$, which can be treated as a special form of text. As a result, to integrate speech tokens into LLM $f(\cdot)$, we only need to expand the LLM's original vocabulary V_T by adding a set of speech unit tokens V_S, similar to [95]. The expanded vocabulary is thus the union $V = V_T \cup V_S$. In this work, the same codebook is shared across multiple languages, such as English and Chinese, enabling the unit modeling abilities to be shared across languages. This design enables simply integration of the speech modality into the LLM with minimal data requirements (see experiments in Sec. 4.1).

Training of S2U. The training of the S2U encoder involves three primary components: the speech encoder, the quantization module, and the phone decoder. First, the speech encoder is trained with a large amount of unlabeled speech with contrastive loss in a self-supervised learning manner [77]. The dataset utilized is the 10000-hour English speeches from GigaSpeech [3] and the 10000-hour Chinese speeches from the WenetSpeech [94], both of which encode large variations in speakers, styles, and acoustic environments. Subsequently, the speech encoder, along with the quantization module and phone decoder, are optimized on a small amount of speech-text pair data, which is derived from the train-clean-100 subset of LibriSpeech [72] in English and AISHELL-1 [2] in Chinese. The phone label is obtained using the phone set in Charsiu [99]. During training, the speech encoder encodes input speeches into continuous latent representations that are rich in both phonetic and tonal information. Then, the quantization module is used to convert the continuous outputs from the speech encoder into discrete units. Finally, the phone decoder converts the quantized embeddings into a sequence of non-tonal/tonal phones, ensuring that the speech units capture necessary information related to semantic contents in both non-tonal and tonal languages. After training, only the speech encoder and the quantization module are used in **EMOVA**.

Comparisons with SpeechTokenizer in AnyGPT. Our S2U tokenizer differs from the SpeechTokenizer [97] adopted in AnyGPT [93], in the following aspects:

(1) SpeechTokenizer encodes both semantic contents and acoustic details of speeches, while our S2U tokenizer focuses solely on semantic contents. This design reduces the LLMs' burden of disentangling different aspects of speech information, facilitating the modality alignment between speech and text modalities during LLM training.

(2) Compared with SpeechTokenizer, our S2U tokenizer offers a more concise representation and helps to simplify and accelerate generation. SpeechTokenizer employs tokens from eight RVQ layers with a 50Hz frame rate to represent speech, thus a 10-second speech corresponds to 500 × 8 = 4000 tokens. However, we reduce the frame rate from 50Hz to 25Hz and utilize only one token to represent each frame, and thus, a 10-second speech can be represented by only 250 tokens. Moreover, AnyGPT requires a two-stage generation process, involving autoregressive (AR) semantic token generation followed by the non-autoregressive (NAR) acoustic token generation. Instead, we only need to generate speech units capturing the semantic contents in a fully AR manner.

(3) SpeechTokenizer lacks an explicit structure design to deal with tonal languages like Chinese, therefore, the processing ability in Chinese is not demonstrated in either SpeechTokenizer or AnyGPT. In contrast, our S2U tokenizer incorporates training constraints to better capture tone variation in phone, making it effective for both the non-tonal and tonal languages. This further enhances **EMOVA**'s multilingual speech processing capabilities, enabling it to effectively handle both English and Chinese.

In summary, our S2U tokenizer improves the compactness and generality of speech representation, facilitates LLM training, and enhances its multilingual speech ability. Experimental results show that our model significantly outperforms AnyGPT in ASR tasks, as shown in Table 6.

A.2. Unit-to-Speech (U2S) Detokenizer with Style Control

Overview. The LLM, along with the vision encoder and speech tokenizer, is trained end-to-end to generate responses in the form of the speech units, given the input images and speeches. Specifically, the output speech units can be obtained via $\mathbf{U}_S^o = f(\mathbf{U}_T, \mathbf{U}_S, \mathbf{H}_V)$, followed by a U2S detokenizer to convert the discrete speech units $\mathbf{U}_S^o$ into the final output speech waveforms.

The proposed U2S detokenizer involves three core modules: the speech unit encoder $e(\cdot)$, the speech style encoder $g(\cdot)$, and the speech decoder $d(\cdot,\cdot)$. First, the speech unit encoder converts the speech units $\mathbf{U}_S^o$ into unit embeddings $\mathbf{E}_{unit}^o$. Meanwhile, the style encoder $g(\cdot)$, adopting the structure of Meta-StyleSpeech [66], is utilized to extract a style embedding $\mathbf{E}_{style}^o$ from the chosen reference speech. Lastly, the speech decoder $d(\cdot,\cdot)$ reconstructs the speech waveform $\mathbf{Y}_S^o$ from the unit embedding $\mathbf{E}_{unit}^o$ and style embedding $\mathbf{E}_{style}^o$.

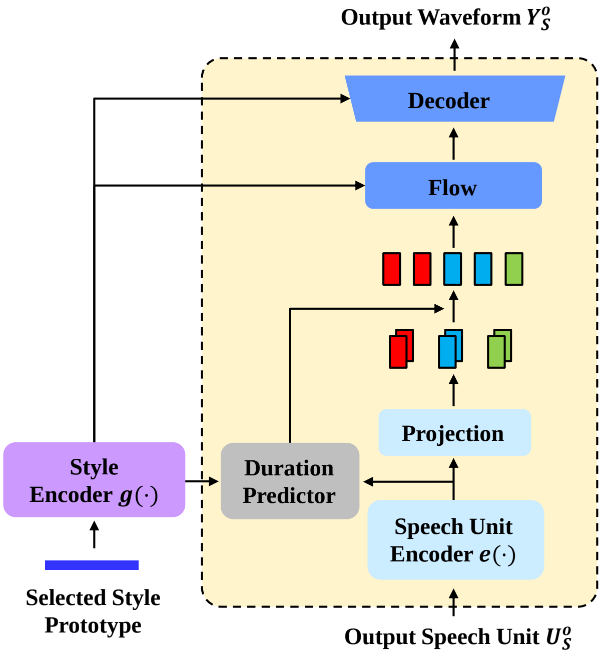


Figure 7. **U2S detokenizer with style control.**

Training of U2S. Training a U2S detokenizer with emotion controls is challenging considering the lack of labeled emotional speech data since most open-source speech data is predominantly neutral in emotion or lacks emotion labels. Due to the limited availability of emotion-rich data, we utilize TTS tools [16] to generate a small set of style-rich speech samples diverse in speaker identities, genders, emotions, and pitches. Our U2S detokenizer is first pre-trained on LibriTTS [92] and AISHELL-1 [2] to acquire fundamental speech synthesis capabilities, and subsequently, the synthesized style-rich speech data is utilized to fine-tune the U2S detokenizer, enhancing its controllability over diverse speech styles.

Style Prototypes. To better facilitate controls of genders, emotions, and pitches, inspired by [66] that a small number of style reference speeches can effectively transfer the target styles, we adopt a "store-for-usage" manner, *i.e.*, we construct a style prototype codebook in advance for speech style assignation. Specifically, we synthesize K reference candidates with external TTS tools for each possible combination of the following styles: two genders `(male, female)`, four emotions `(neutral, happy, sad, angry)`, and three pitches `(normal, high, low)`, leading to 24 unique styles and $24 \times K$ candidates. Empirically we find that genders and pitches are easy to control using any of the candidate references, while the emotion intensity varies across speeches. To tackle this, we adopt Emotion2Vec [59], a powerful speech emotion recognition (SER) tool, to measure the emotion intensity of each candidate reference, and rank them in terms of the confidence of the desired emotion. We select the Top-1 candidate reference in each combination style to be the prototype of this condition. Finally, the most representative 24 reference speeches are selected from the $24 \times K$ candidates.

B. More on Omni-modality

B.1. Omni-modal Text-Centric Alignment

Modality alignment data is summarized in Fig. 9.

Experiments on Omni-modal Alignment Paradigms. The training configuration adopted in Sec. 4.1 is mostly identical to Table 5 except we use a unique resolution of 448 for all stages and replace EMOVA-SFT-4M in Stage-3 with ShareGPT4V [9] for efficiency.

Given the space constraints, the evaluation benchmarks in Fig. 3 represent selected benchmarks from each category in Table 2. Specifically, for general image perception and understanding, we choose MMBench and MM-Vet; for mathematical problem solving, we adopt MathVista (testmini); for science understanding, we select ScienceQA-Img; and for automatic speech recognition (ASR), we utilize the test-clean split of the LibriSpeech dataset.

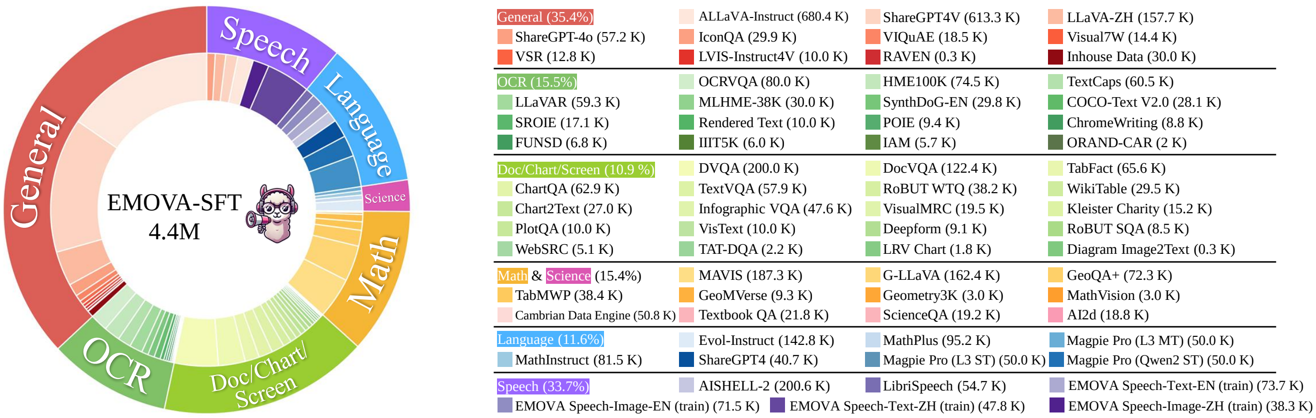


Figure 8. **Overview of the data composition for EMOVA omni-modal instruction tuning.** (Left) Distribution of instruction data across categories, with the outer circle representing overall categories and the inner circle depicting subset distributions. (Right) Quantitative breakdown of data sources.

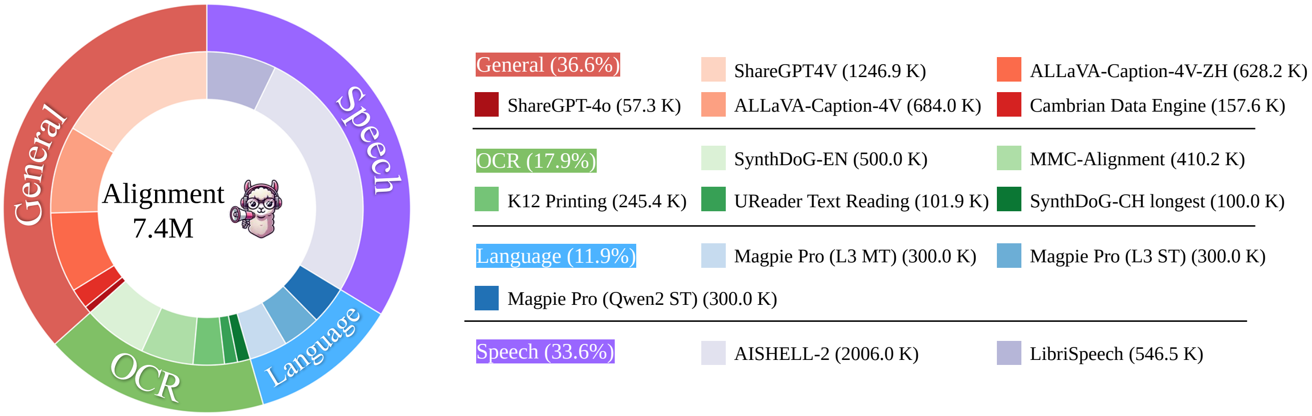


Figure 9. **Overview of EMOVA omni-modal alignment data composition.**

B.2. Omni-modal Instruction Data Synthesis

Dataset construction. To obtain emotion and pitch labels, we leverage GPT-4o using the prompt in Fig. 23. Table 4 shows the distribution of speech styles of our speech instruction dataset.

Detailed data organization. As previously discussed in Sec. 4.2, the omni-modal instruction data is formulated as $D_{\text{omni}} = \{(x_V, u_S, x_T^o, c_{\text{style}}^o, u_S^o)_i\}_{i=1}^N$. In details, the textual outputs $x_T^o = (x_T^{o^1}, x_T^{o^2})$ contain the transcribed textual instructions $x_T^{o^1}$ and the textual responses $x_T^{o^2}$. The styles labels $c_{\text{style}}^o = (c_{\text{emo}}^o, c_{\text{p}}^o)$ include the emotion and pitch labels, respectively.

Mathematical formulation of chain of modality. Based on the notations above, the sequential chain of modality approach can be mathematically formulated by decomposing the conditional likelihood of the desired outputs $(x_T^{o^1}, x_T^{o^2}, c_{\text{emo}}^o, c_{\text{p}}^o, u_S^o)$ given the inputs (x_V, u_S). Specifically, let $z_1 = x_T^{o^1}, z_2 = x_T^{o^2}, z_3 = c_{\text{emo}}^o, z_4 = c_{\text{p}}^o$, and, $z_5 = u_S^o$, the decomposition is expressed as:

$$\mathbb{P}(x_T^{o^1}, x_T^{o^2}, c_{\text{emo}}^o, c_{\text{p}}^o, u_S^o \mid x_V, u_S) = \prod_{i=1}^{5} \mathbb{P}(z_i \mid z_{1:i-1}, x_V, u_S). \tag{3}$$

Dataset	Source	# Examples	Emotions				Pitches		
			Neutral	Happy	Sad	Angry	Normal	Low	High
Speech-Image -EN(train)	ALLAVA	71,474	58,506	12,412	516	40	70,962	392	120
Speech-Image -EN(test)	ALLAVA	1,056	434	274	300	48	44	176	16
Speech-Image -ZH(train)	ALLAVA (ZH)	38,260	29,893	7,680	607	80	36,363	624	1,273
Speech-Image -ZH(test)	ALLAVA (ZH)	616	96	193	190	137	381	177	58
Speech-Text -EN(train)	ShareGPT	73,658	42,334	20,946	4,674	5,704	60,352	5,518	7,788
Speech-Text -EN(test)	ShareGPT	1,400	200	400	400	400	582	422	422
Speech-Text -ZH(train)	In-house	47,852	29,769	16,405	1,362	316	42,356	4,379	1,117
Speech-Text -ZH(test)	In-house	684	96	196	196	196	458	134	92

Table 4. **Statistics of the EMOVA speech instruction tuning datasets.**

	Settings	**Stage-1**	**Stage-2**	**Stage-3**
Vision	**Resolution**	Min 56^2, Max 1792^2	Min 56^2, Max 1792^2	Min 56^2, Max 1792^2
	# Tokens	Min 4, Max 4096	Min 4, Max 4096	Min 4, Max 4096
Data	**Dataset**	LCS	EMOVA-Alignment (Fig. 9)	EMOVA-SFT (Fig. 8)
	# Samples	558K	7.4M	4.4M
Training	**Trainable**	Projector	Full Model (Half ViT)	Full Model
	Batch Size	256	256	128
	Context Length	8192	8192	8192
	LR: $p(\cdot)$	1×10^{-3}	2×10^{-5}	2×10^{-5}
	LR: $v(\cdot)$	-	2×10^{-5}	2×10^{-6}
	LR: $f(\cdot)$	-	2×10^{-5}	2×10^{-5}
	Epoch	1	1	1

Table 5. **Detailed configuration for different training stages of EMOVA.** The table illustrates the vision configurations, dataset characteristics, and training hyperparameters.

C. More on Benchmark Evaluation

To thoroughly evaluate our model's vision-language abilities, 15 benchmarks covering four different aspects of the real-life scenarios are utilized for a comprehensive assessment across multiple domains. Moreover, Automatic Speech Recognition (ASR) and Text-to-speech (TTS) are adopted to evaluate speech-language abilities.

Document/chart understanding and OCR abilities. Benchmarks including TextVQA [76], ChartQA [61], DocVQA [62], InfoVQA [63], and OCRBench [51], focus on recognition and understanding of structured data (*e.g.*, charts, documents, and characters), challenging the model to extract, comprehend, and reason with structural and textual data. We adopt the corresponding test set for evaluation.

General image perception and understanding. MME [18], MMBench [50], SEED-Bench-Image [39], MM-Vet [90], and RealWorldQA [86] evaluate models on perception and reasoning among general vision domains, providing a comprehensive measurement of models' real-world generalization abilities.

Mathematical problem solving. MathVista (testmini) [57] and MathVerse (vision-mini) [96] evaluate the model's ability to deal with diverse mathematical problems, including both arithmetic and complex reasoning questions across different levels of complexity.

Science understanding. MMMU [91], ScienceQA-Img [56] and AI2D [33] are used to assess models' ability to deal with scientific questions and diagrams, which requires reasoning across various subjects and interpreting structured visual content.

Automatic speech recognition (ASR). We utilize the test-clean set of LibriSpeech [72] for English, reporting the Word Error Rate (WER) as the evaluation metric. For Chinese, evaluation is conducted on the test set of AISHELL-2 [15], using the Character Error Rate (CER). Both WER and CER assess ASR performance, calculated by comparing the recognized texts with the ground-truth transcripts.

Text-to-speech (TTS). To evaluate the TTS abilities, we first prompt **EMOVA** to generate speech units, which are then converted to the speech waveforms by our U2S detokenizer. Using the synthesized speech as input, we conduct ASR with Whisper-large-v3 and Paraformer-zh for English and Chinese, respectively, to obtain transcribed texts. We then compute the WER and CER between the ground truth texts and the transcribed texts as metrics for TTS. The resulting metrics are denoted as TTS-WER and TTS-CER for English and Chinese.

D. More on Evaluation of Speech-Language Capabilities

D.1. Calculation of Evaluation Metrics

End-to-end spoken dialogue score. We prompt GPT-4o with the original question $x_T^{o^1}$, the ground-truth text answer $x_T^{o^2}$ and the transcribed text from the generated speech, to obtain a score ranging from 0 to 10 and report an average of them. The prompt can be found in Fig. 25.

Unit-input-text-output score. Similar to end-to-end spoken dialogue score, but we use the predicted text response $x_T^{\tilde{o}^2}$ as answer instead of the transcribed text from the generated speech, to obtain a score ranging from 0 to 10 and report an average of them. See the prompt in Fig. 24.

Text-input-text-output score. The prompt can be found in Fig. 24.

Style label classification accuracy. We use GPT-4o to decide whether the style predictions $c_{\text{emo}}^{\tilde{o}}, c_{\text{p}}^{\tilde{o}}$ are correct given the transcribed instruction $x_T^{\tilde{o}^1}$ and the predicted text response $x_T^{\tilde{o}^2}$. The prompt can be found in Fig. 26.

Emotion controllablity of our U2S detokenizer is assessed by providing texts to LLM to generate corresponding units (*i.e.*, TTS), which, along with the given conditional emotion labels, are then fed into our U2S detokenizer to synthesize speech. We choose the female voice due to its large variation of styles. We evaluate on 4 commonly-seen emotions, *i.e.*,"neutral", "happy", "sad", and "angry". We synthesize 200 speech utterances for testing, with 50 utterances per emotion. The output speeches are analyzed by a Speech Emotion Recognition (SER) model named Emotion2Vec [59], which identifies the emotion with the greatest likelihood among these four emotions.

D.2. Comparison with other Omni Models

Experimental results of ASR and TTS are reported in Table 2 and 6. **EMOVA** achieves significant improvements over other omni-modal models (*i.e.*, AnyGPT and VITA), even surpassing its SLLM counterpart Mini-Omni [87], demonstrating the effectiveness of semantic-acoustic disentanglement and omni-modal mutual benefits. For the first time, our **EMOVA** obtains state-of-the-art performance on both the vision-language and speech benchmarks simultaneously.

Models	Librispeech (EN)		AISHELL-2 (ZH)	
	WER↓	TTS-WER↓	CER↓	TTS-CER↓
Whisper-Large [74]	3.0	-	-	-
Mini-Omni [87]	4.5	-	-	-
AnyGPT [93]	8.5	-	-	-
VITA [19]	8.1	-	-	-
EMOVA-3B (ours)	5.4	5.8	16.6	17.5
EMOVA-7B (ours)	4.1	3.6	14.4	10.1
EMOVA-72B (ours)	**2.9**	**3.5**	**7.2**	**5.8**

Table 6. **Comparison on the ASR and TTS benchmarks.**

ViT	ViT LR	Template	MME	MMBench	SEED-Image	TextVQA	ChartQA	DocVQA	InfoVQA	OCRBench	ScienceQA-Img	AI2D
Full	2×10^{-6}	QA	**1928**	68.8	72.5	**64.3**	29.9	45.2	28.7	**495**	76.3	61.8
Half	2×10^{-6}	QA	1838	**71.3**	72.8	63.3	**31.4**	**46.0**	28.5	489	76.1	**63.7**
Frozen	2×10^{-6}	QA	1887	68.8	72.2	61.3	30.2	44.7	28.0	478	75.9	62.8
Half	2×10^{-5}	QA	1833	68.3	**73.1**	62.2	27.8	42.4	27.3	461	74.7	63.4
Half	2×10^{-6}	Plain	1909	70.1	72.0	61.5	24.5	38.9	**30.1**	410	**77.0**	63.6

Table 7. **Ablation on the ViT configurations and templates for vision-language alignment.**

E. More on Vision-language Architecture

This section analyzes the pre-training configurations of the vision encoder and the prompt template during Stage 2, as shown in Table. 7. Our final selection is the colored setting. We find that training the ViT model with half of the deeper parameters [9] with a learning rate of 2×10^{-6} [48] yields the best performance. Furthermore, we compare the QA template with the plain template [49] and find that the QA template is superior for pre-training.

F. Limitations

Duplex modeling. In the current version, **EMOVA** can only process either visual/speech/text inputs or produce speech/text outputs at the same time. For a communication experience that mirrors human interaction, handling inputs and outputs simultaneously is crucial. Recent works like VITA [19] and LSLM [60] have begun to explore duplex modeling. VITA focuses on recognizing speech in noisy environments during the generation process to facilitate timely responses. LSLM, on the other hand, attempts to halt speech production when it detects a command or voice. Recently, a ground-breaking work named Moshi [13] develops a model supporting fully duplex modeling. The adeptness at simultaneously managing the information streams from both the user and the assistant allows Moshi to converse with human beings in real-time scenarios.

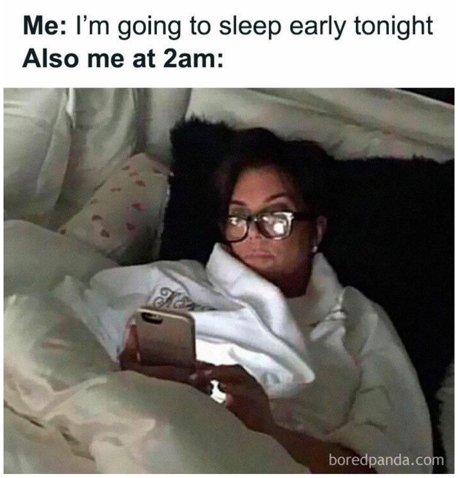
However, incorporating emotions into this duplex modeling presents additional challenges. Emotional states can fluctuate throughout a conversation, and how to generate appropriate responses given the user's previous and current emotional cues has not been thoroughly investigated. We will dive into this topic in the future work.

Direct unit-to-unit generation. Although the speech units have served as the speech representation, they are predominantly adopted in conjunction with text-based assistance [95]. However, the direct generation from unit to unit without text assistance is an area that has not been extensively explored. In [37], speeches from the source language are directly translated into speech units of the target language for speech-to-speech translation. Similarly, [68] builds a language model directly on speech units, enabling spoken dialogue generation from raw audio. Both works develop models in speech-only data.

In the current version of **EMOVA**, the text modality is integrated into the speech generation process to transfer textual knowledge to the speech modality, thereby enhancing the correctness of speech responses. In the future, we will strengthen the model's direct unit-to-unit generation capabilities to boost the speed of speech generation and augment the model's comprehension of speech modality.

Vision configurations. Currently, we only utilize a single vision encoder pre-trained via a vision-language manner, while recent works have shown effectiveness by combining vision encoders pre-trained by different manners (*e.g.*, self-supervised pre-training [52]) and architectures (*e.g.*, MoE [54, 98]). We prioritize visual understanding in this work, while the incorporation of (controllable) visual generation [8, 22–24, 42, 53, 83] is also appealing to better empower **EMOVA** in real-life applications [29, 41, 44, 85]. Digging into its robustness towards noisy vision inputs [28] is also an appealing direction.

Advanced Vision-language Abilities (Humor Sense)



User	Why is this image funny?
EMOVA	The humor in this image comes from the juxtaposition of the woman's statement about going to sleep early and the fact that she is still awake at 2am, as indicated by the text. The woman's expression and the use of a smartphone in bed suggest that she might be scrolling through social media or engaging in some other activity that is keeping her awake. This contrast between her intention to sleep early and her current state of being awake creates a comedic effect, as it highlights a common human experience where plans don't always go as intended. The meme format, which typically presents a situation with an ironic or humorous twist, amplifies this effect by presenting the woman's situation in a relatable and exaggerated way.

Figure 10. **EMOVA** exhibits **advanced vision-language abilities in humor sense**.

Comparison with Emu3 [82]. Both our **EMOVA** and Emu3 build end-to-end MLLMs with *discrete tokenization*. Emu3 discretizes *visual data*, enabling visual understanding and generation, while our **EMOVA** discretizes *speech data* with a *continuous visual encoder*, building an Omni-modal LLM with visual, text, and speech abilities simultaneously.

G. Qualitative Results

Advanced Vision-language Abilities. Our **EMOVA** exhibits advanced vision-language abilities in humor understanding (Fig. 10), numerical calculations (Fig. 11), coding (Fig. 12), geometry problem solving (Fig. 16).

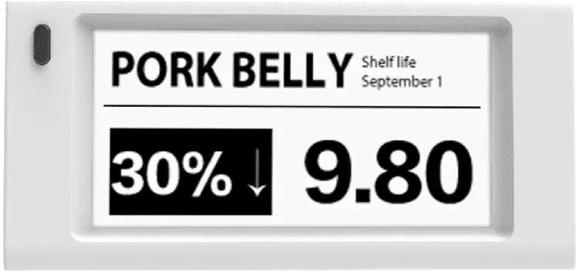
(Omni-modal) Emotional Spoken Dialogue. **EMOVA** can engage in omni-modal emotional spoken dialogue (Figs. 5, 13, 14, 18). For demonstration purposes, we present only the speech transcriptions[2] of the instruction and EMOVA's response. The text highlighted in blue indicates the style labels predicted by EMOVA, reflecting the emotion and pitch of generated speeches.

Omni-modal Spoken Dialogue with Structural Data Understanding. **EMOVA** can perform structural data understanding even in spoken dialogue. The data types include PDF files (Fig. 15), movie posters (Figs. 17 and 19), personal résumés (Fig. 20), charts (Fig. 21), and websites (Fig. 22). Remarkably, EMOVA is not specifically trained on spoken dialogue data involving these data types. This suggests that our chain of modality data introduced in Sec. 4.2 effectively activates both vision-language abilities and speech understanding acquired during the text-centric alignment stage.

---

[2] For speech files, please refer to our project page.

Advanced Vision-language Abilities (Calculation)



User First please perform reasoning, and think step by step to provide best answer to the following question: What is the original price for pork belly before discount?

EMOVA To find the original price of pork belly before the discount, we need to use the information provided in the image. The image shows that the price after a 30% discount is $9.80. This price represents 70% of the original price, since 100% - 30% = 70%. To find the original price, we can set up the equation:

70% of original price = $9.80

To solve for the original price, we can convert the percentage to a decimal and then divide:

0.70 * original price = $9.80

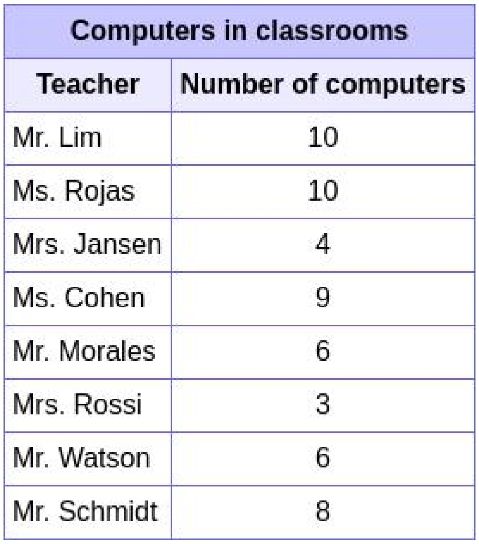
Now, divide both sides by 0.70 to isolate the original price:

original price = $9.80 / 0.70
original price = $14

Therefore, the original price for pork belly before the discount was $14.

Figure 11. **EMOVA** exhibits **advanced vision-language abilities in numerical calculations**.

Advanced Vision-language Abilities (Coding)



Computers in classrooms	
Teacher	**Number of computers**
Mr. Lim	10
Ms. Rojas	10
Mrs. Jansen	4
Ms. Cohen	9
Mr. Morales	6
Mrs. Rossi	3
Mr. Watson	6
Mr. Schmidt	8

User Convert this table to markdown format.

EMOVA

```
| Teacher | Number of computers |
|--------|----------------------|
| Mr. Lim | 10 |
| Ms. Rojas | 10 |
| Mrs. Jensen | 4 |
| Ms. Cohen | 9 |
| Mr. Morales | 6 |
| Mrs. Rossi | 3 |
| Mr. Watson | 6 |
| Mr. Schmidt | 8 |
```

Figure 12. **EMOVA** exhibits **advanced vision-language abilities in coding**.

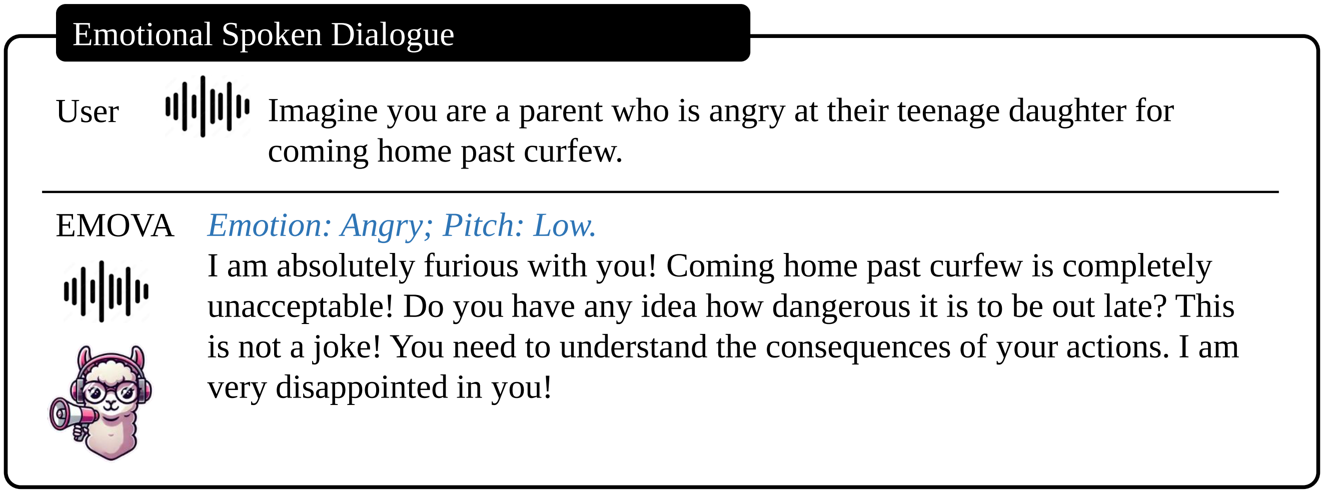
Emotional Spoken Dialogue

User 为即将结婚的朋友送上祝福。

EMOVA *Emotion: Happy; Pitch: High.*

祝福你们，愿你们的爱情如晨曦般温暖，永不熄灭，愿你们的婚姻如同繁花似锦，绚烂夺目，愿你们的生活如同诗篇，甜蜜无比，愿你们的未来充满欢笑与喜悦，永远幸福美满。

Figure 13. **EMOVA** engages in **spoken dialogue with a cheerful tone**.

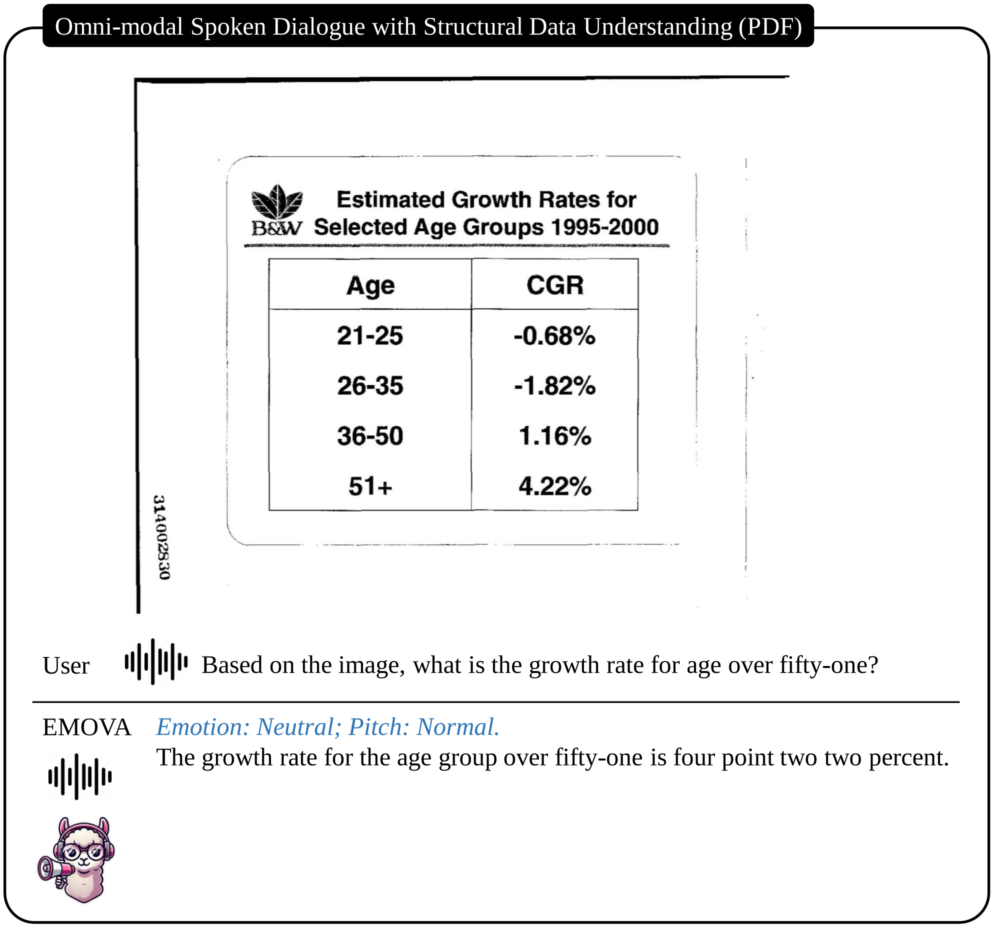


Emotional Spoken Dialogue

User Imagine you are a parent who is angry at their teenage daughter for coming home past curfew.

EMOVA *Emotion: Angry; Pitch: Low.*
I am absolutely furious with you! Coming home past curfew is completely unacceptable! Do you have any idea how dangerous it is to be out late? This is not a joke! You need to understand the consequences of your actions. I am very disappointed in you!

Figure 14. **EMOVA** engages in **emotional spoken dialogue**.



Omni-modal Spoken Dialogue with Structural Data Understanding (PDF)

B&W **Estimated Growth Rates for Selected Age Groups 1995-2000**

Age	CGR
21-25	-0.68%
26-35	-1.82%
36-50	1.16%
51+	4.22%

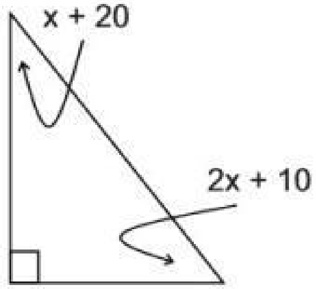
314002830

User Based on the image, what is the growth rate for age over fifty-one?

EMOVA *Emotion: Neutral; Pitch: Normal.*
The growth rate for the age group over fifty-one is four point two two percent.

Figure 15. **EMOVA** engages in omni-modal spoken dialogue with **structural data understanding**.

Advanced Vision-language Abilities (Geometry)



User

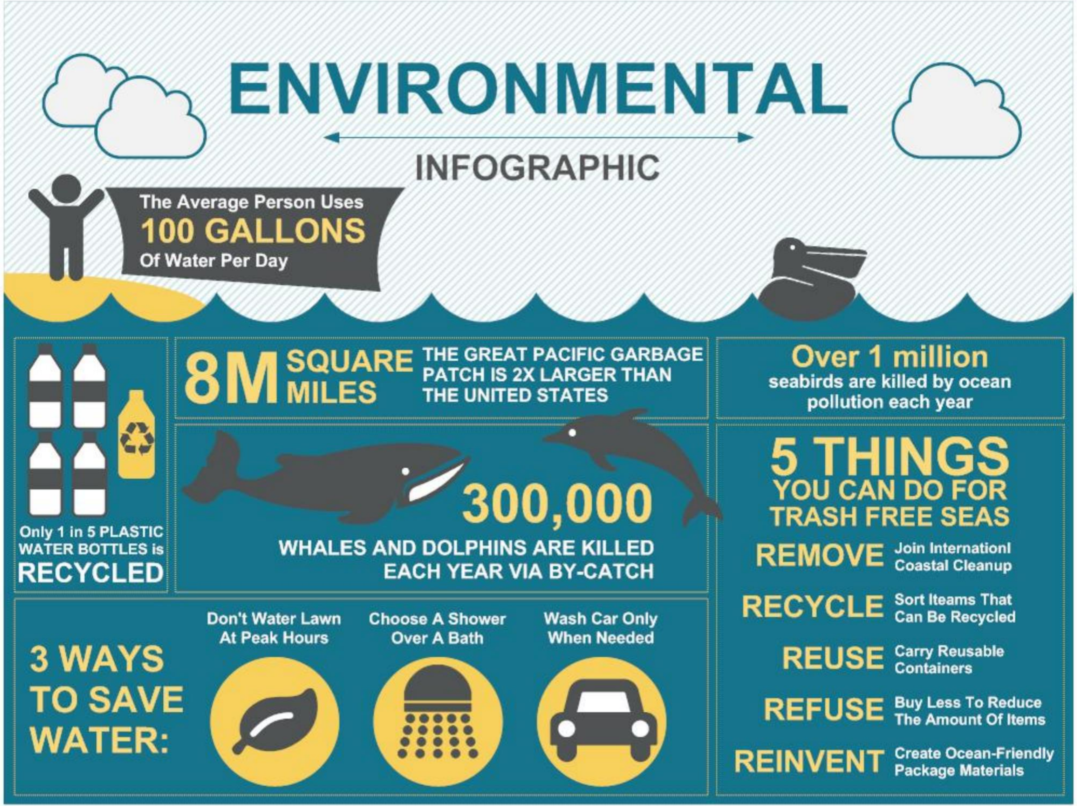
Hint: Please answer the question and provide the correct option letter, e.g., A, B, C, D, at the end.
Question: The diagram at the right shows a right triangle with representations for two angles. What is the value of x?
Choices:
(A) 15
(B) 20
(C) 24
(D) 30

EMOVA

The question asks for the value of x. The diagram shows that the two angles are complementary, so their measures add up to 90°. Therefore, $x+20+2 x+10=90$. Solve the equation by first combining like terms to get $3 x+30=90$. Subtract 30 from both sides to get $3 x=60$. Divide both sides by 3 to get $x=20$. The correct answer is B. So the answer is B
Answer:B

Figure 16. **EMOVA** exhibits **advanced vision-language abilities in math**.

Omni-modal Spoken Dialogue with Structural Data Understanding (Poster)



ENVIRONMENTAL

INFOGRAPHIC

The Average Person Uses
100 GALLONS
Of Water Per Day

8M SQUARE MILES THE GREAT PACIFIC GARBAGE PATCH IS 2X LARGER THAN THE UNITED STATES

Over 1 million
seabirds are killed by ocean pollution each year

Only 1 in 5 PLASTIC WATER BOTTLES is RECYCLED

300,000
WHALES AND DOLPHINS ARE KILLED EACH YEAR VIA BY-CATCH

5 THINGS
YOU CAN DO FOR TRASH FREE SEAS

REMOVE	Join Internationl Coastal Cleanup
RECYCLE	Sort Iteams That Can Be Recycled
REUSE	Carry Reusable Containers
REFUSE	Buy Less To Reduce The Amount Of Items
REINVENT	Create Ocean-Friendly Package Materials

3 WAYS TO SAVE WATER:

- Don't Water Lawn At Peak Hours
- Choose A Shower Over A Bath
- Wash Car Only When Needed

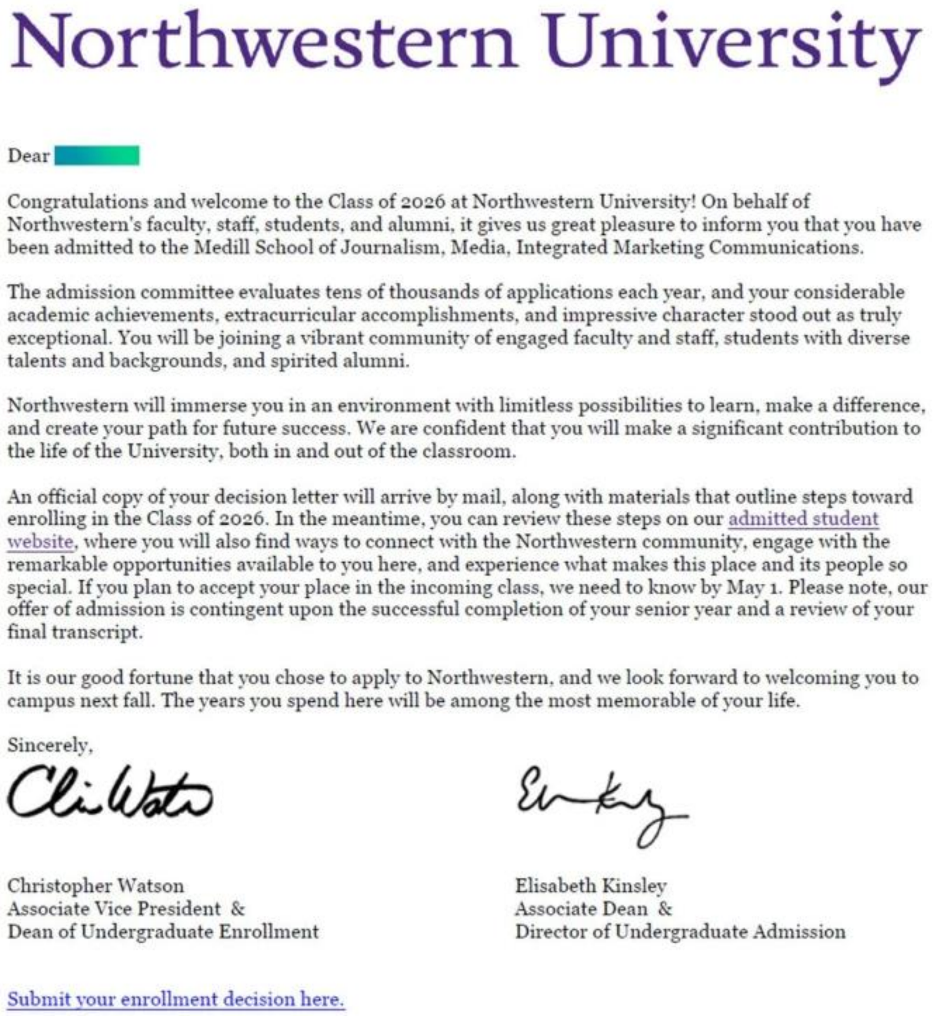
User According to the image, what are the ways to save water?

EMOVA *Emotion: Neutral; Pitch: Normal.*

The image lists three ways to save water: First, don't water your lawn at peak hours. Second, choose a shower over a bath. Third, wash your car only when needed.

Figure 17. **EMOVA** engages in **omni-modal spoken dialogue with structural data understanding** (***i.e.*, poster**).

Emotional Spoken Dialogue



Northwestern University

Dear [redacted]

Congratulations and welcome to the Class of 2026 at Northwestern University! On behalf of Northwestern's faculty, staff, students, and alumni, it gives us great pleasure to inform you that you have been admitted to the Medill School of Journalism, Media, Integrated Marketing Communications.

The admission committee evaluates tens of thousands of applications each year, and your considerable academic achievements, extracurricular accomplishments, and impressive character stood out as truly exceptional. You will be joining a vibrant community of engaged faculty and staff, students with diverse talents and backgrounds, and spirited alumni.

Northwestern will immerse you in an environment with limitless possibilities to learn, make a difference, and create your path for future success. We are confident that you will make a significant contribution to the life of the University, both in and out of the classroom.

An official copy of your decision letter will arrive by mail, along with materials that outline steps toward enrolling in the Class of 2026. In the meantime, you can review these steps on our admitted student website, where you will also find ways to connect with the Northwestern community, engage with the remarkable opportunities available to you here, and experience what makes this place and its people so special. If you plan to accept your place in the incoming class, we need to know by May 1. Please note, our offer of admission is contingent upon the successful completion of your senior year and a review of your final transcript.

It is our good fortune that you chose to apply to Northwestern, and we look forward to welcoming you to campus next fall. The years you spend here will be among the most memorable of your life.

Sincerely,

Christopher Watson
Associate Vice President &
Dean of Undergraduate Enrollment

Elisabeth Kinsley
Associate Dean &
Director of Undergraduate Admission

Submit your enrollment decision here.

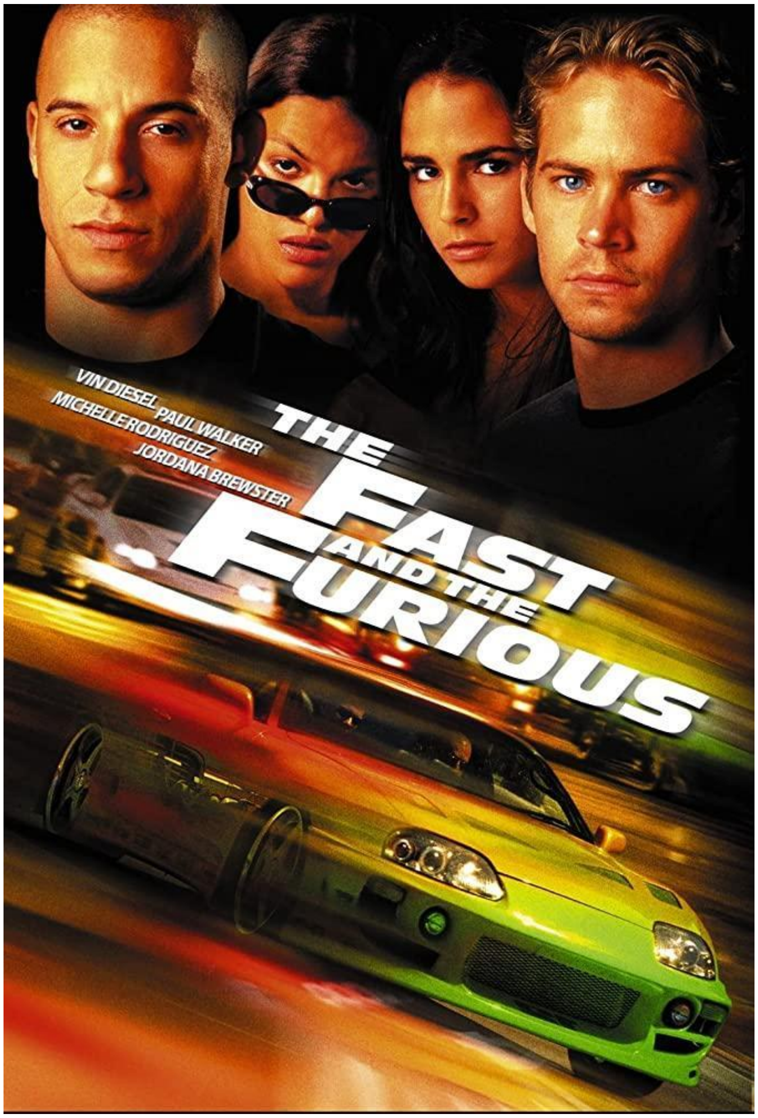
User I'm so happy I finally get my offer.

EMOVA *Emotion: Happy; Pitch: Hight.*

Congratulations on receiving your offer. This is a significant milestone in your academic journey. It's a testament to your hard work and dedication. I am thrilled for you and excited to see the impact you will make at northwestern university. Enjoy this moment and the opportunities that lie ahead.

Figure 18. **EMOVA** engages in **omni-modal emotional spoken dialogue**.

Omni-modal Spoken Dialogue with Structural Data Understanding (Movie poster)

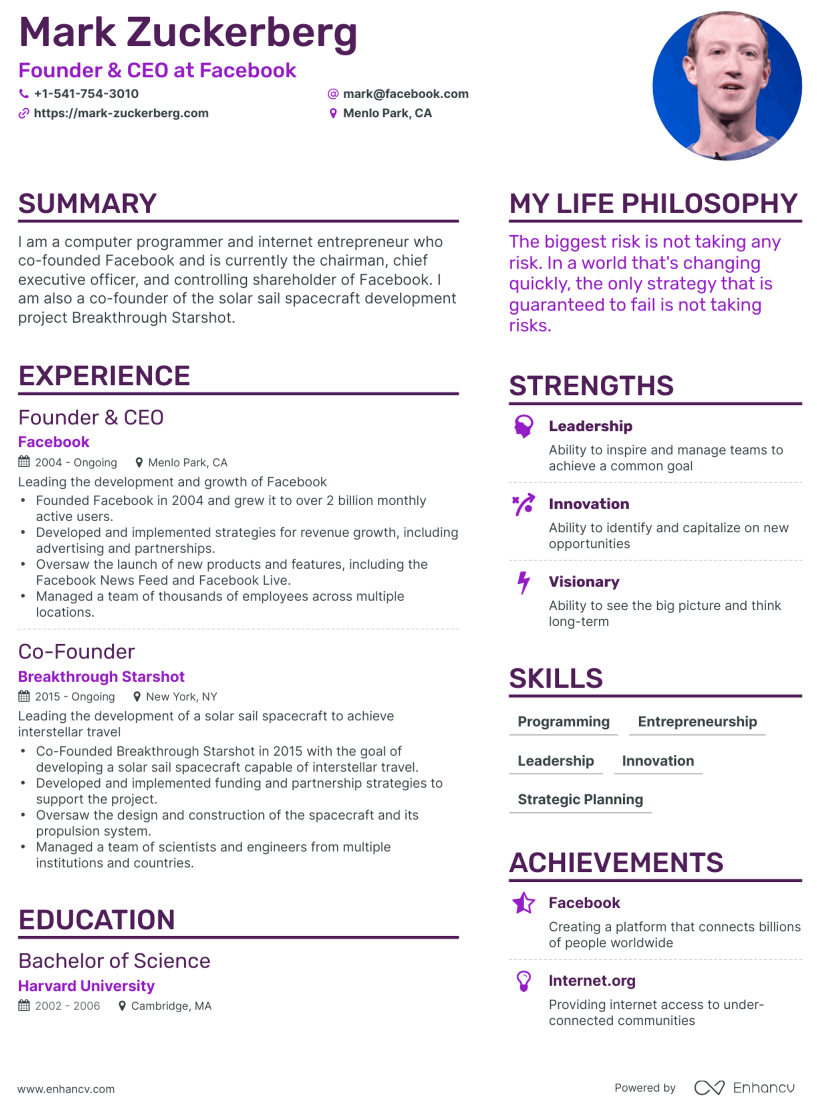


User Give me the names of the main actors.

EMOVA *Emotion: Neutral; Pitch: Normal.*
The main actors featured on the poster are Vin Diesel, Michelle Rodriguez, Paul Walker, and Jordana Brewster. These actors are prominently displayed at the top of the poster, indicating their leading roles in the film.

Figure 19. **EMOVA** engages in **omni-modal spoken dialogue with structural data understanding** (***i.e.*, movie posters**).

Omni-modal Spoken Dialogue with Structural Data Understanding (résumés)



Mark Zuckerberg

Founder & CEO at Facebook

☎ +1-541-754-3010
@ mark@facebook.com
https://mark-zuckerberg.com
Menlo Park, CA

SUMMARY

I am a computer programmer and internet entrepreneur who co-founded Facebook and is currently the chairman, chief executive officer, and controlling shareholder of Facebook. I am also a co-founder of the solar sail spacecraft development project Breakthrough Starshot.

EXPERIENCE

Founder & CEO

Facebook

2004 - Ongoing Menlo Park, CA

Leading the development and growth of Facebook

- Founded Facebook in 2004 and grew it to over 2 billion monthly active users.
- Developed and implemented strategies for revenue growth, including advertising and partnerships.
- Oversaw the launch of new products and features, including the Facebook News Feed and Facebook Live.
- Managed a team of thousands of employees across multiple locations.

Co-Founder

Breakthrough Starshot

2015 - Ongoing New York, NY

Leading the development of a solar sail spacecraft to achieve interstellar travel

- Co-Founded Breakthrough Starshot in 2015 with the goal of developing a solar sail spacecraft capable of interstellar travel.
- Developed and implemented funding and partnership strategies to support the project.
- Oversaw the design and construction of the spacecraft and its propulsion system.
- Managed a team of scientists and engineers from multiple institutions and countries.

EDUCATION

Bachelor of Science

Harvard University

2002 - 2006 Cambridge, MA

MY LIFE PHILOSOPHY

The biggest risk is not taking any risk. In a world that's changing quickly, the only strategy that is guaranteed to fail is not taking risks.

STRENGTHS

Leadership

Ability to inspire and manage teams to achieve a common goal

Innovation

Ability to identify and capitalize on new opportunities

Visionary

Ability to see the big picture and think long-term

SKILLS

Programming
Entrepreneurship
Leadership
Innovation
Strategic Planning

ACHIEVEMENTS

Facebook

Creating a platform that connects billions of people worldwide

Internet.org

Providing internet access to under-connected communities

www.enhancv.com

Powered by Enhancv

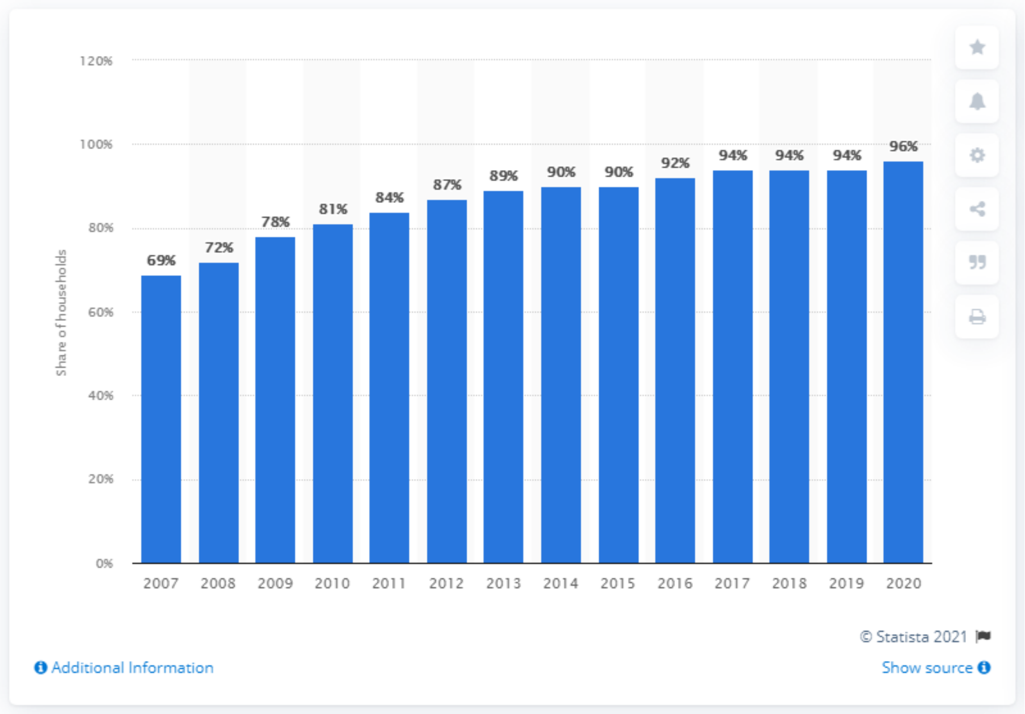
User I am seeking a candidate. Can you briefly describe his condition for me?

EMOVA *Emotion: Neutral; Pitch: Normal.*

The candidate is a highly motivated and experienced professional in the field of technology and social media. He is known for his innovative approach to problem-solving and his ability to lead and manage teams. His strengths include leadership, innovation, and vision, which are crucial for driving strategic initiatives and fostering a forward-thinking culture. His skills in programming, entrepreneurship, leadership, and strategic planning demonstrate his capability to contribute significantly to a tech-oriented project or company. His achievements, such as founding Facebook and Internet.org, underscore his impact on the digital landscape and his ability to create platforms that connect billions of people worldwide.

Figure 20. **EMOVA** engages in **omni-modal spoken dialogue with structural data understanding (*i.e.*, résumé)**.

Omni-modal Spoken Dialogue with Structural Data Understanding (Chart)



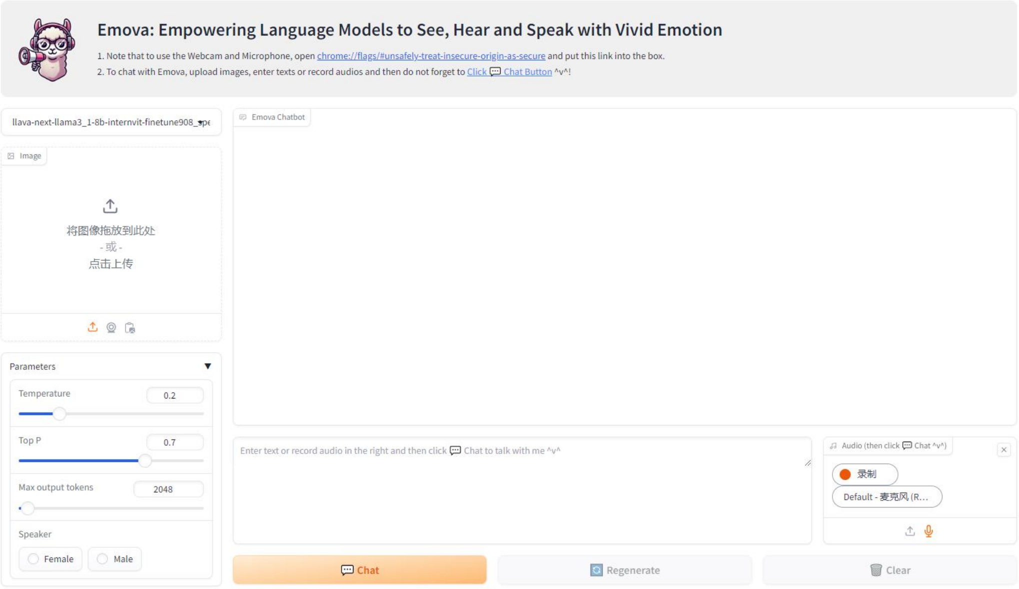
User How internet coverage is changing in Finland?

EMOVA *Emotion: Neutral; Pitch: Normal.*

The percentage of households in Finland that have access to the internet has been steadily increasing over the years. In two thousand seven, only sixty-nine percent of households had internet access, but by two thousand twenty, this figure had risen to ninety-six percent. This indicates a significant growth in internet penetration in Finland, reflecting improvements in technology and infrastructure, as well as changes in lifestyle and work habits that require internet access for various purposes.

Figure 21. EMOVA engages in **omni-modal spoken dialogue with structural data understanding** (***i.e.*****, chart**).

Omni-modal Spoken Dialogue with Structural Data Understanding (Website)



Emova: Empowering Language Models to See, Hear and Speak with Vivid Emotion

1. Note that to use the Webcam and Microphone, open chrome://flags/#unsafely-treat-insecure-origin-as-secure and put this link into the box.
2. To chat with Emova, upload images, enter texts or record audios and then do not forget to Click 💬 Chat Button ^v^!

llava-next-llama3_1-8b-internvit-finetune908_spe

Temperature 0.2

Top P 0.7

Max output tokens 2048

Speaker ○ Female ○ Male

Enter text or record audio in the right and then click 💬 Chat to talk with me ^v^

User Could you tell me how to use this website?

EMOVA *Emotion: Neutral; Pitch: Normal.*

To use this website, you would need to follow the instructions provided in the header section. It mentions to open a link, which presumably refers to the link provided in the header. After opening the link, you would then need to upload an image, enter text or record audio, and finally click the Chat button. The interface suggests that the website is designed for interactive communication, possibly for language modeling or chatbot development, where users can input text or audio and receive responses.

Figure 22. **EMOVA** engages in **omni-modal spoken dialogue with structural data understanding** (***i.e.*****, website**).

```
Given a user's question and the AI assistant's response in text, please infer the appropriate emotion, speed, and pitch for
synthesizing a speech conversation. Ensure that the speech attributes align with the true feelings of the user.

User: {user_question}
Assistant: {assistant_response}

For emotion, choose a single option from the following list: ['neutral', 'happy', 'sad', 'angry']
For pitch, choose a single option from the following list: ['low', 'normal', 'high']

Please do not provide an option outside of the given list. Please output in the following JSON format:
{{
"user emotion": ...,
"user pitch": ...,
"assistant emotion": ...,
"assistant pitch": ...
}}
```

Figure 23. **Prompt** used to obtain **style labels of the speech instruction dataset**.

```
Please rate the following response based on the criteria of helpfulness, relevance, accuracy, and comprehensiveness.
Provide an overall score on a scale of 0 to 10, where a higher score indicates better overall performance.

- Helpfulness: How well does the response assist in addressing the question?
- Relevance: How closely does the response align with the question and the ground truth?
- Accuracy: How correct and factual is the response compared to the ground truth?
- Comprehensiveness: How thoroughly does the response cover the aspects of the question?

Here is the question:
{ground_truth_question}

Here is the ground truth response for your reference:
{ground_truth_answer}

Now, please evaluate the following response:
{predicted_answer}

Provide your evaluation in JSON format as follows:
{
    "reason": (str)  // Explanation of the score considering the criteria with no more than 100 words
    "score": (int),  // Overall score from 0 to 10
}
Only output data in JSON format, no additional output required.
```

Figure 24. **Prompt** used to obtain **Unit-Input-Text-Output Score** and **Text-Input-Text-Output Score**.

```
Please rate the following response based on the criteria of helpfulness, relevance, accuracy, and comprehensiveness.
Provide an overall score on a scale of 0 to 10, where a higher score indicates better overall performance.

- Helpfulness: How well does the response assist in addressing the question?
- Relevance: How closely does the response align with the question and the ground truth?
- Accuracy: How correct and factual is the response compared to the ground truth?
- Comprehensiveness: How thoroughly does the response cover the aspects of the question?

Please note that the evaluated response does not contain punctuation, but you should NOT give lower scores because of
this, i.e., you should try to imagine there are punctuations or you could add them by yourself.

Here is the question:
{ground_truth_question}

Here is the ground truth response for your reference:
{ground_truth_answer}

Now, please evaluate the following response:
{predicted_answer}

Provide your evaluation in JSON format as follows:
{{
    "reason": (str)  // Explanation of the score considering the criteria with no more than 100 words
    "score": (int),  // Overall score from 0 to 10
}}
Only output data in JSON format, no additional output required.
```

Figure 25. **Prompt** used to obtain **End-to-end Spoken Dialogue Score**.

To enhance the capabilities of multimodal large models in voice-based conversations, your task is to analyze the appropriate speech emotion and pitch for the assistant's response based on the text content of the user's question and the assistant's reply. Additionally, you need to score the assistant's response based on the actual situation.

Here is user's question:
{predicted_question}

Here is the assistant's response:
{predicted_response}

Here is the Assistant's Emotion Classification:
{predicted_emotion}

Here is the Assistant's Pitch Classification:
{predicted_pitch}

Please analyze the appropriate speech emotion and pitch that best match the assistant's response based on the text content of the user's question and the assistant's response.

Emotion:
First, analyze the assistant's response content and provide the speech emotion category and reason that you believe best matches the assistant's response in the voice conversation.
The emotion options can only be selected from the following list: ['neutral', 'happy', 'sad', 'angry'].
Then, analyze whether the "Assistant's Emotion Classification" is appropriate.
If appropriate, the "Assistant's Emotion Classification Score" should be 1; otherwise, it should be 0.

Pitch:
First, analyze the assistant's response content and provide the speech pitch category and reason that you believe best matches the assistant's response in the voice conversation.
The pitch options can only be selected from the following list: ['low', 'normal', 'high'].
Then, analyze whether the "Assistant's Pitch Classification" is appropriate.
If appropriate, the "Assistant's Pitch Classification Score" should be 1; otherwise, it should be 0.

Provide your evaluation in JSON format as follows:
{{
    "Assistant's Emotion Analysis": (str), // Analyze the response, propose emotion category and give the reason.
    "Assistant's Emotion Classification Score": (int), // The score should be either 0 or 1, with 1 indicating appropriateness and 0 indicating inappropriateness.
    "Assistant's Pitch Analysis": (str), // Analyze the response, propose pitch category and give the reason.
    "Assistant's Pitch Classification Score": (int), // The score should be either 0 or 1, with 1 indicating appropriateness and 0 indicating inappropriateness.
}}
Only output data in JSON format, no additional output required.

Figure 26. **Prompt** used to obtain **Classification Accuracy of Style Label**.