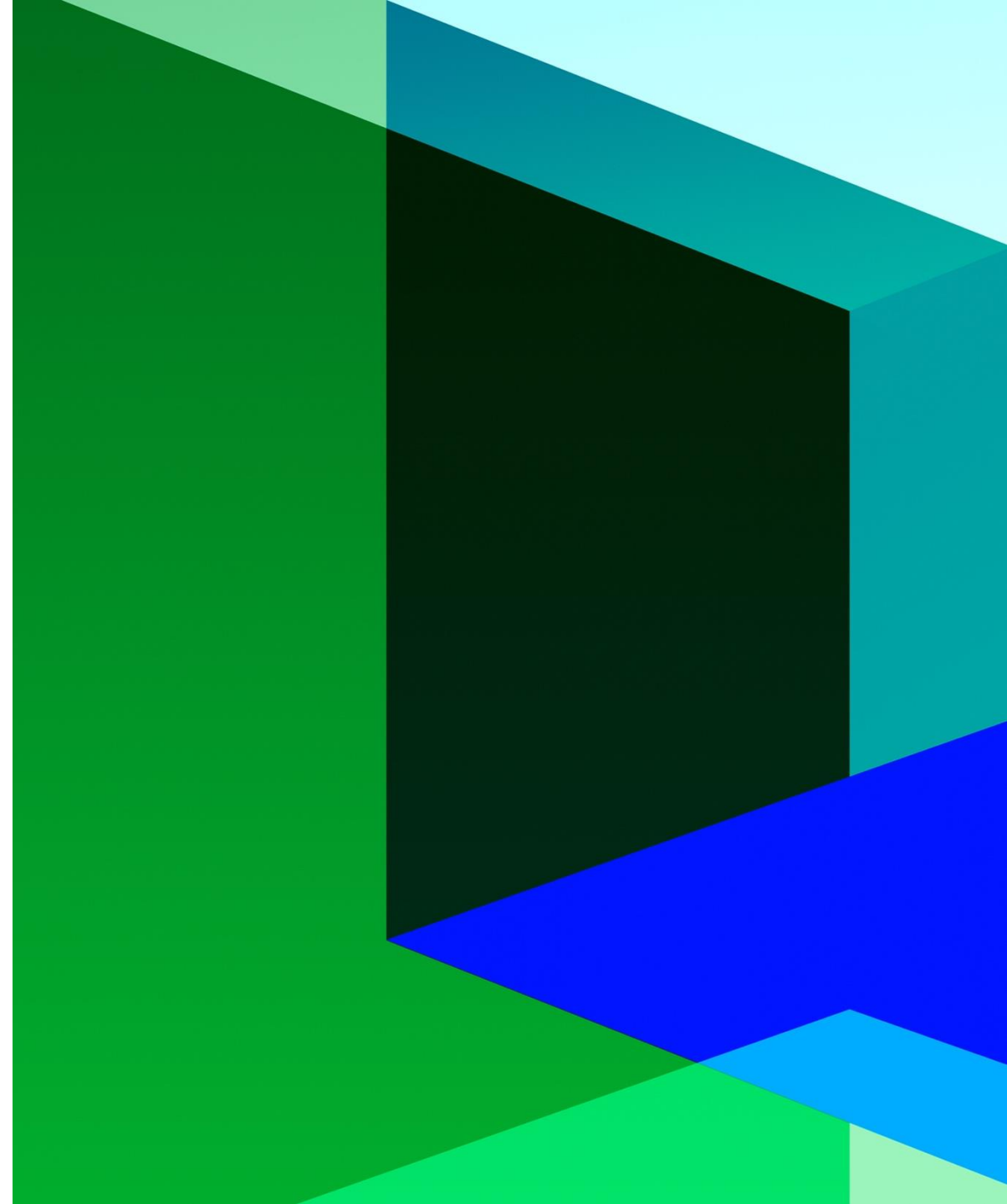


Building Lightweight Multimodal Models for Enterprise AI

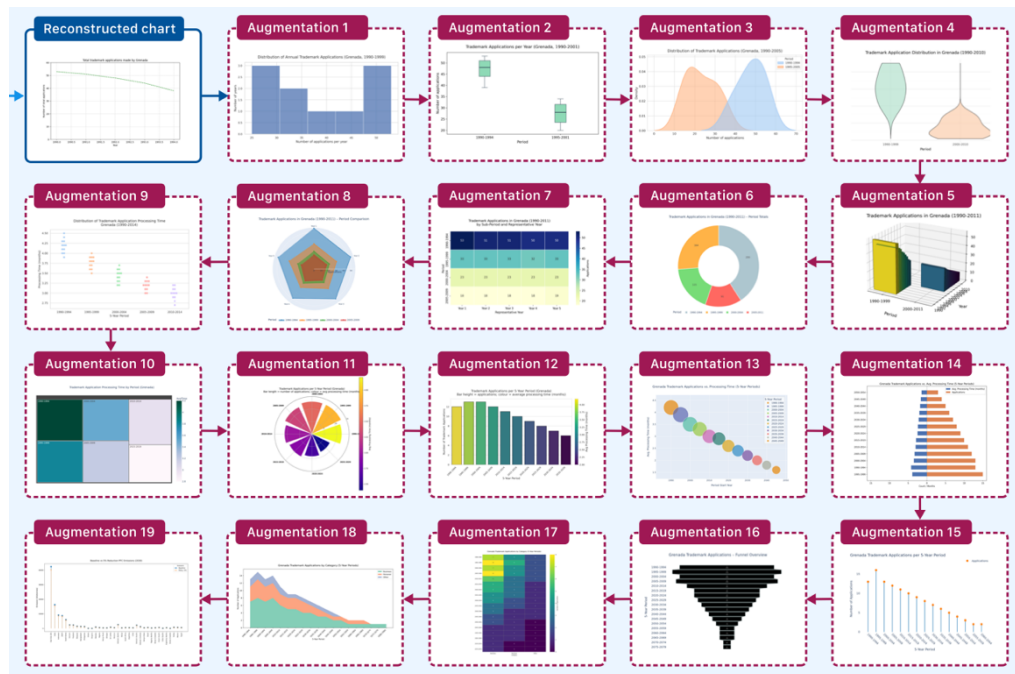
January 21, 2026



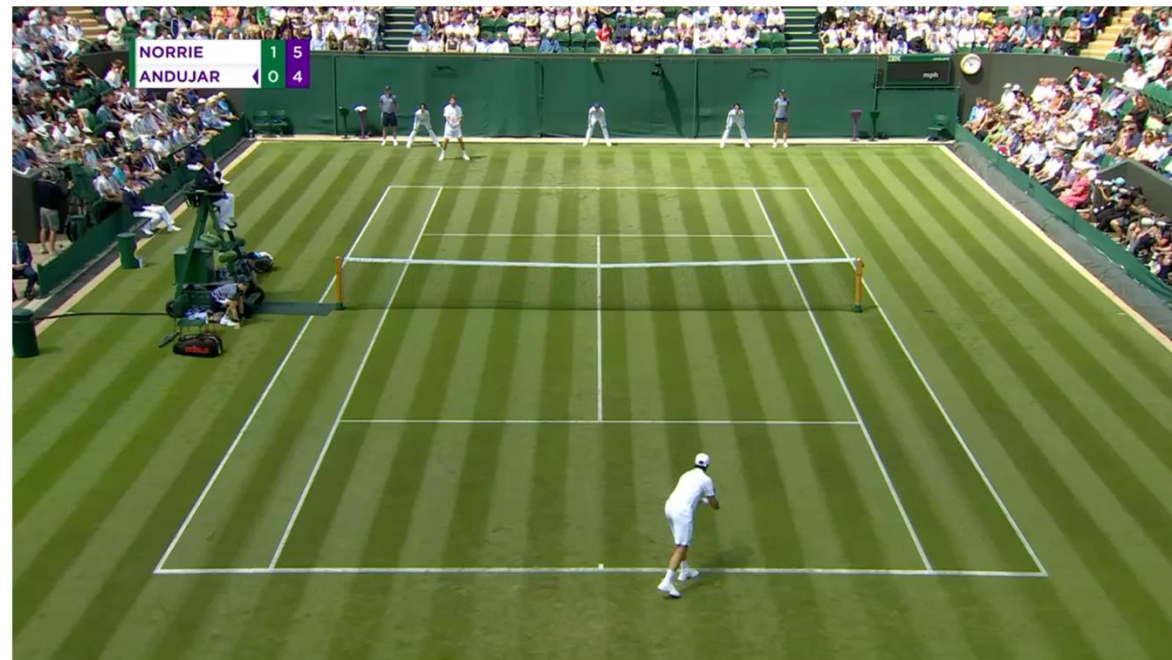
MIT-IBM Computing Research Lab



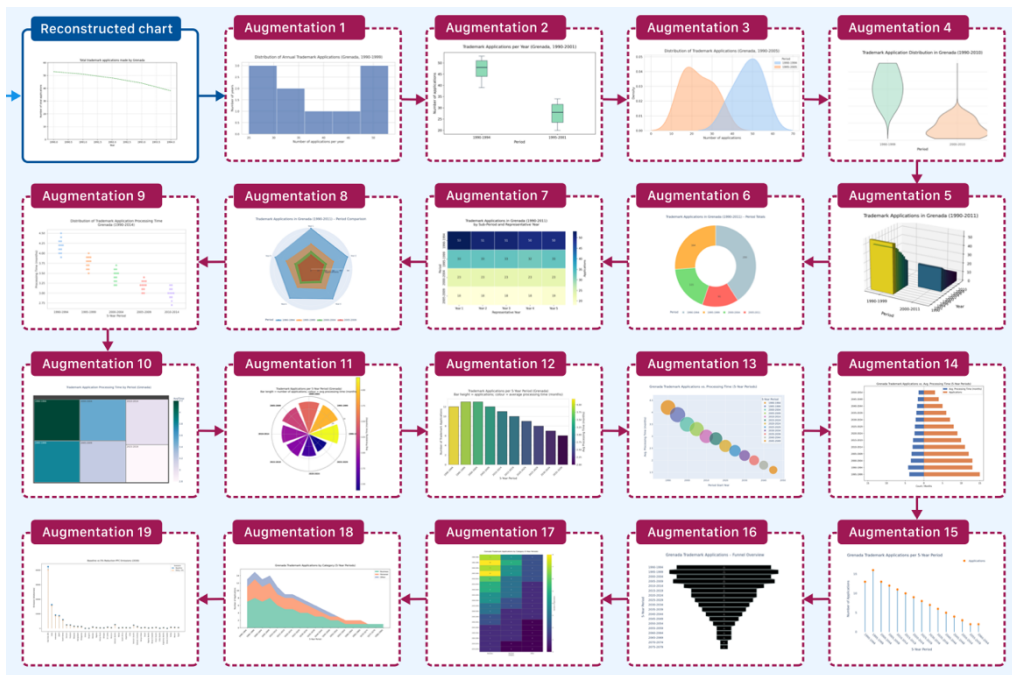
1. Granite Vision



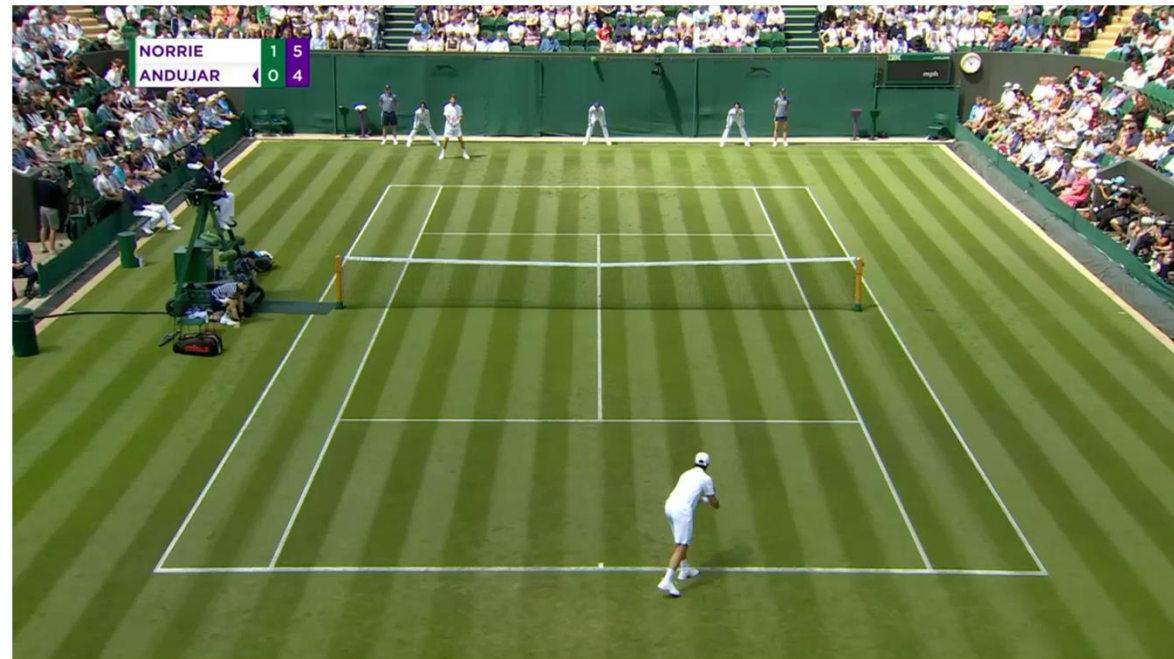
2. AI Commentary



1. Granite Vision



2. AI Commentary



3. Research Outlook: Self-Evolving AI Memory

Granite Vision Team

IBM Cambridge, Haifa, Zurich, Yorktown, Almaden

arXiv > cs > arXiv:2502.09927

Search...

Help | Advan

Computer Science > Computer Vision and Pattern Recognition

[Submitted on 14 Feb 2025]

Granite Vision: a lightweight, open-source multimodal model for enterprise Intelligence

Granite Vision Team: Leonid Karlinsky, Assaf Arbelle, Abraham Daniels, Ahmed Nassar, Amit Alfassi, Bo Wu, Eli Schwartz, Dhiraj Joshi, Jovana Kondic, Nimrod Shabtay, Pengyuan Li, Roei Herzig, Shafiq Abedin, Shaked Perek, Sivan Harary, Udi Barzelay, Adi Raz Goldfarb, Aude Oliva, Ben Wieles, Bishwaranjan Bhattacharjee, Brandon Huang, Christoph Auer, Dan Gutfreund, David Beymer, David Wood, Hilde Kuehne, Jacob Hansen, Joseph Shtok, Ken Wong, Luis Angel Bathen, Mayank Mishra, Maksym Lysak, Michele Dolfi, Mikhail Yurochkin, Nikolaos Livathinos, Nimrod Harel, Ophir Azulai, Oshri Naparstek, Rafael Teixeira de Lima, Rameswar Panda, Sivan Doveh, Shubham Gupta, Subhro Das, Syed Zawad, Yusik Kim, Zexue He, Alexander Brooks, Gabe Goodhart, Anita Govindjee, Derek Leist, Ibrahim Ibrahim, Aya Soffer, David Cox, Kate Soule, Luis Lastras, Nirmitt Desai, Shila Ofek-koifman, Sriram Raghavan, Tanveer Syeda-Mahmood, Peter Staar, Tal Drory, Rogerio Feris

We introduce Granite Vision, a lightweight large language model with vision capabilities, specifically designed to excel in enterprise use cases, particularly in visual document understanding. Our model is trained on a comprehensive instruction-following dataset, including document-related tasks, such as content extraction from tables, charts, diagrams, sketches, and infographics, as well as general image tasks. The architecture of Granite Vision is centered around visual modality alignment with a decoder-only, 2 billion parameter Granite large language model. Additionally, we introduce a dedicated safety classification approach in test-time that leverages a sparse set of attention vectors to identify potential harmful inputs. Despite its lightweight architecture, Granite Vision achieves strong results in standard benchmarks related to visual document understanding, as well as on the LiveXiv benchmark, which is designed to avoid test set contamination by using a constantly updated corpus of recently published Arxiv papers. We are releasing the model under the Apache-2 license, allowing for both research and commercial use, while offering complete visibility into the training data and other relevant details. See [this https URL](#) for model weights.

Subjects: **Computer Vision and Pattern Recognition (cs.CV)**; Artificial Intelligence (cs.AI)

Cite as: [arXiv:2502.09927](#) [cs.CV]

(or [arXiv:2502.09927v1](#) [cs.CV] for this version)

<https://doi.org/10.48550/arXiv.2502.09927> 

Submission history

From: Rogerio Feris [[view email](#)]

[v1] Fri, 14 Feb 2025 05:36:32 UTC (14,193 KB)

Granite Vision for Enterprise AI

- First version more focused on document images, including plots, charts, flow diagrams, tables, infographics, ...
- Beyond documents: applications in other domains (semiconductors, insurance, UI screens, etc.)

RECEIPT

East Repair Inc.
1912 Harvest Lane
New York, NY 12210

LOGO

BILL TO
John Smith
2 Court Square
New York, NY 12210

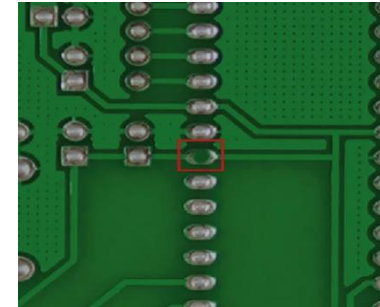
SHIP TO
John Smith
3787 Phenix Drive
Cambridge, MA 12210

RECEIPT #
P.O.#
DUE DATE

US-001
11/02/2019
23/12/2019
26/02/2019

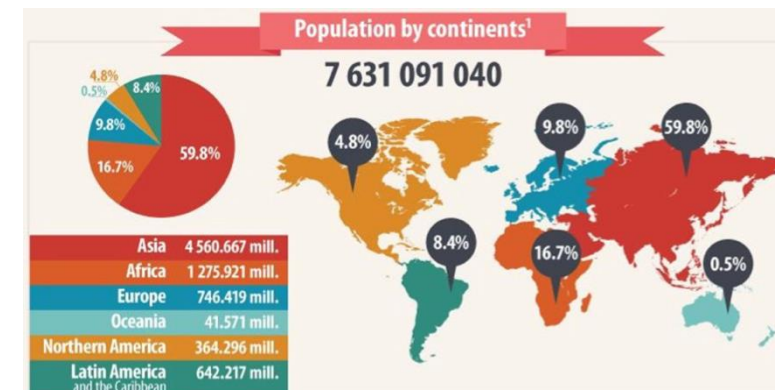
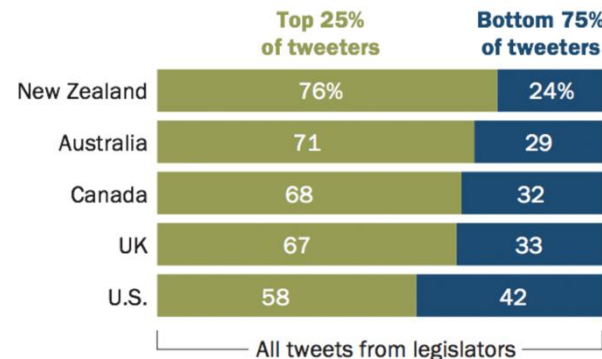
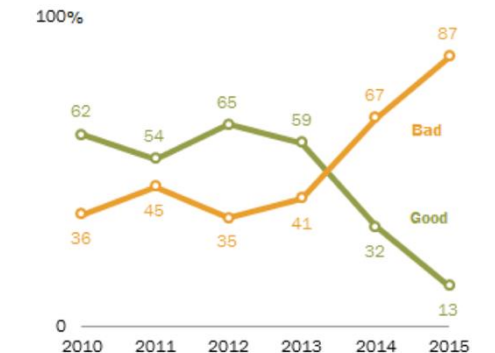
QTY	DESCRIPTION	UNIT PRICE	AMOUNT
1	Front and rear brake cables	100.00	100.00
2	New set of pedal arms	15.00	30.00
3	Labor 3hrs	5.00	15.00
Subtotal			145.00
Sales Tax 6.25%			9.06
TOTAL			\$154.06

John Smith



Rapid Decline in Brazilians' Assessment of Economy

Current economic situation in Brazil is ...





BULL BAR & RESTAURANT
2800S, BULL PARK AVE
NEW HILLS, NC, 27154-0806
919-868-8668

ORDER : 246 12/12/2019
HOST : SUSUIE WAG 02:16 PM

1 WINE BOTTLE	\$ 19.99
2 CHEESE STEAK	\$ 17.90
1 SPICY BARQUE STEAK	\$ 12.49

VISA 0890 SALE

SUBTOTAL \$ 50.46
TAX 2.02
TOTAL: \$ 52.48

TRANSACTION TYPE: SALE
AUTHORIZATION: APPROVED
PAYMENT CODE: 7659054432297
PAYMENT ID: 706995403426400
CARD READER: SWIPED/CHIP

* TIP: -----
=TOTAL: -----

X -----

CUSTOMER COPY
THANKS FOR VISITING
BULL BAR & RESTAURANT

[illegible]

\$52.48

Granite Vision



1 WINE BOTTLE, 2 CHEESE STEAK, 1 SPICY BARQUE STEAK

Granite Vision



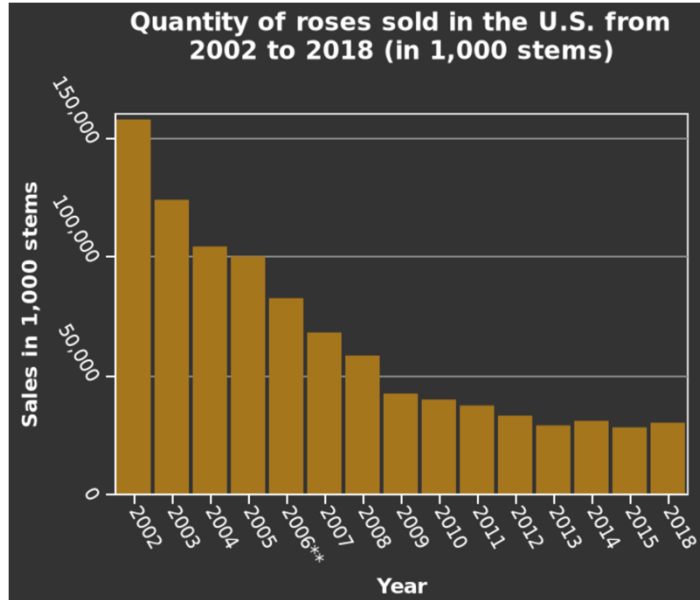
The document is an Interest Accrual Form for the National and Community Service Corporations. It outlines the process for accruing interest on a qualified student loan during a member's term of national service. The form requires the member to provide their personal information, loan details, and certify the accuracy of the information provided. The member must also authorize the release of loan information to the National Service Trust.

Granite Vision

Is this problem solved???

Qwen3-VL

Convert this chart image to CSV format



Year	Sales in 1,000 stems
2002	150,000
2003	105,000
2004	90,000
2005	85,000
2006	70,000
2007	55,000
2008	45,000
2009	40,000
2010	35,000
2011	33,000
2012	31,000
2013	29,000
2014	28,000
2015	27,000
2016	26,000
2017	25,000
2018	24,000

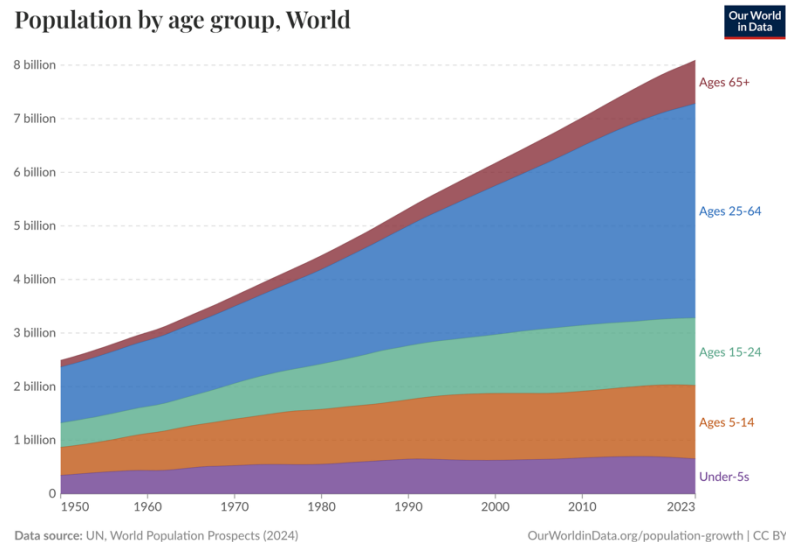


Granite 4.0 Vision

Year	Sales
2002	157,500
2003	124,000
2004	104,500
2005	100,500
2006**	82,500
2007	68,500
2008	58,500
2009	42,500
2010	40,000
2011	37,500
2012	33,500
2013	29,000
2014	31,000
2015	28,500
2018	30,000



Qwen3-VL



"This is a stacked area chart titled \"Population by age group, World,\" which illustrates the global population distribution across different age groups from 1950 to 2023. The chart is sourced from the UN's World Population Prospects (2024) and is presented by Our World in Data. The horizontal axis represents the years, spanning from 1950 to 2023. The vertical axis represents the population in billions, ranging from 0 to 8 billion. The chart is divided into five distinct age groups, each represented by a different color: Under-5s: Represented by a purple area. Ages 5-14: Represented by an orange area. Ages 15-24: Represented by a light green area. Ages 25-64: Represented by a blue area. Ages 65+: Represented by a dark red area. The data shows a clear demographic shift over time: In 1950, the population was approximately 1 billion for the Under-5s, 1 billion for Ages 5-14, 2 billion for Ages 15-24, 3 billion for Ages 25-64, and 4 billion for Ages 65+. By 2023, the population has grown significantly. The Under-5s group has increased to about 1.1 billion, Ages 5-14 to 1.2 billion, Ages 15-24 to 1.5 billion, Ages 25-64 to 2.5 billion, and the oldest group, Ages 65+, has reached approximately 7.5 billion. The chart demonstrates that the proportion of the population aged 65 and over has been growing steadily, while the proportion of the working-age population (Ages 25-64) has also increased. The total population, which has grown from about 2.5 billion in 1950 to over 8 billion in 2023, reflects the trend of an aging global population."

Granite 4.0 Vision

"The chart is titled \"Population by age group, World\" and is an area chart that visually represents the population distribution across different age groups over time. The x-axis represents the years, ranging from 1950 to 2023, while the y-axis represents the population in billions. The chart is divided into five distinct age groups, each represented by a different color: Under-5s in purple, Ages 5-14 in orange, Ages 15-24 in green, Ages 25-64 in blue, and Ages 65+ in brown. The data shows a clear trend of population growth across all age groups over the years. In 1950, the population for each age group was relatively small, with Ages 25-64 having the largest share at approximately 1.1 billion, followed by Ages 5-14 at around 0.6 billion, Ages 15-24 at about 0.5 billion, Under-5s at roughly 0.4 billion, and Ages 65+ at approximately 0.15 billion. By 2023, the population for each age group has significantly increased. Ages 25-64 remains the largest group, reaching around 4.0 billion, followed by Ages 5-14 at about 1.4 billion, Ages 15-24 at approximately 1.3 billion, Under-5s at roughly 0.75 billion, and Ages 65+ at around 0.9 billion. The chart effectively illustrates the aging population trend, with the Ages 25-64 group consistently having the highest population share throughout the years. The Ages 65+ group, while smaller in comparison, shows a notable increase, indicating an aging population. The visual representation with distinct colors for each age group helps in easily comparing the population growth and distribution across different age categories over time."

Positive Feedback

- More than half-million downloads (HuggingFace + Ollama), since February 2025
- Positive feedback from the community, especially given the size of the model (3B)

Jeffrey Morgan, Co-founder and CEO of Ollama



r/ollama • 3 days ago
ML-Future

Ollama with granite3.2-vision is excellent for OCR and for processing text afterwards

granite3.2-vision: I just want to say that after a day of testing it is exactly what I was looking for.

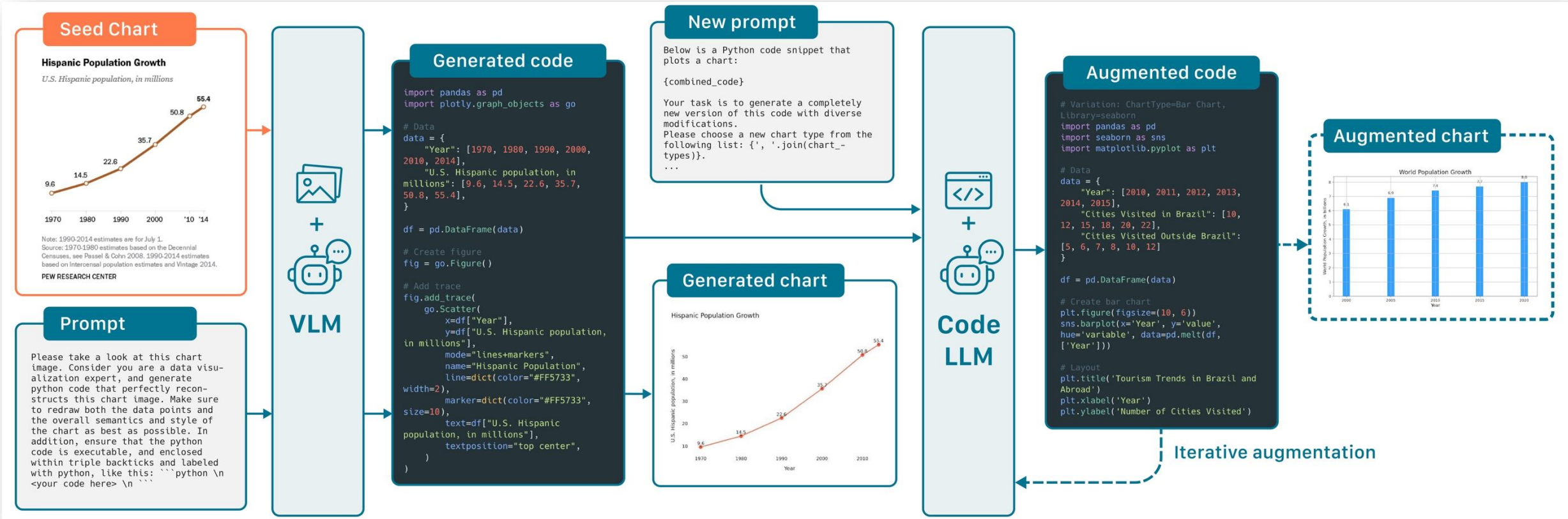
It can work perfectly locally with less than 12gb of ram.

I have tested it to interpret some documents in Spanish and then process their data. Considering its size the performance and precision are surprising.

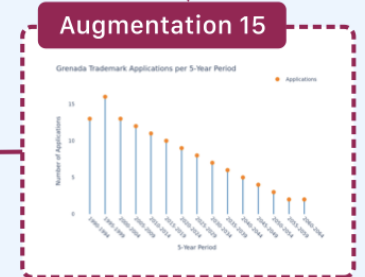
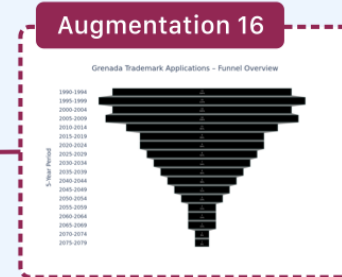
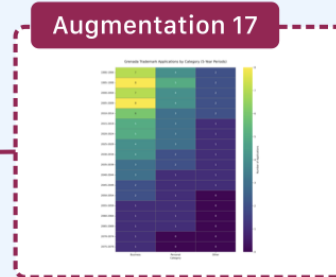
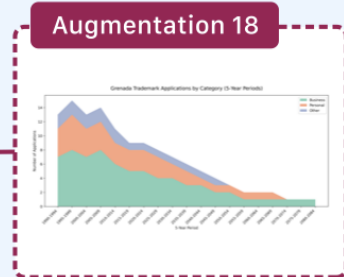
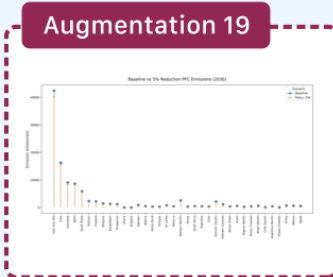
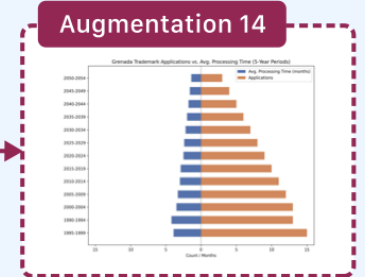
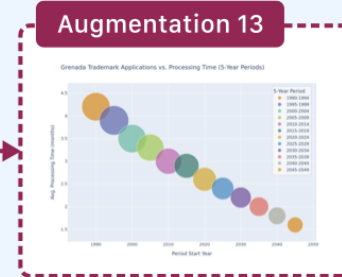
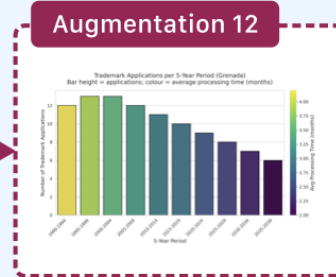
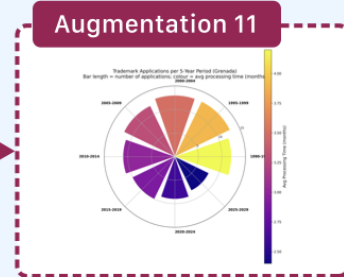
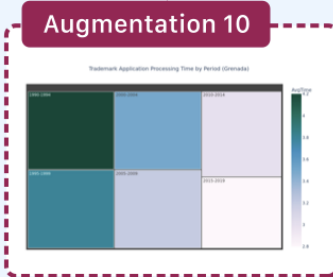
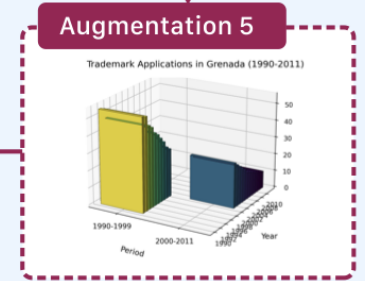
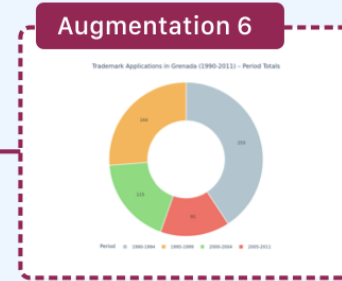
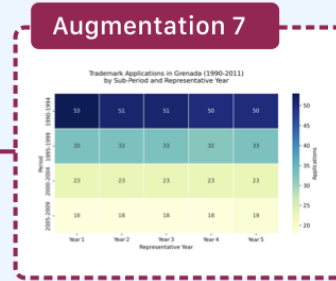
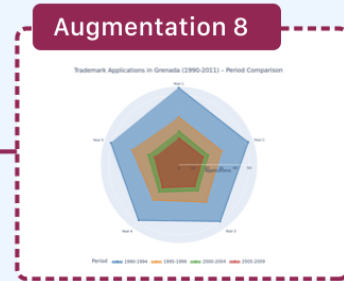
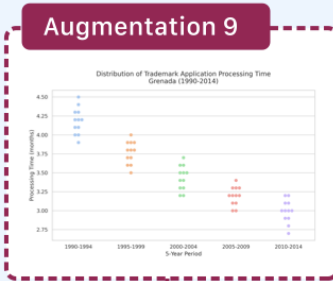
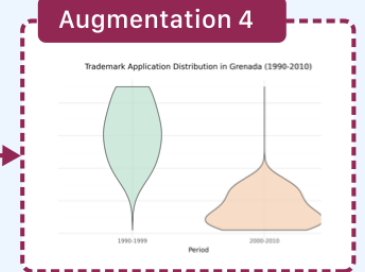
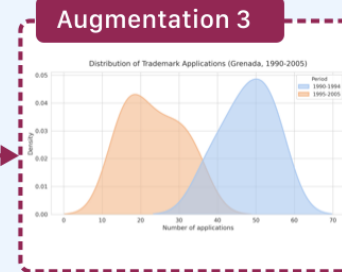
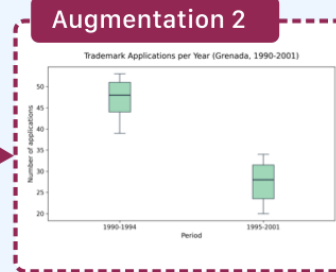
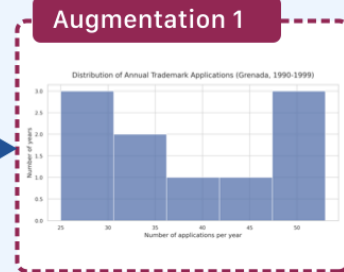
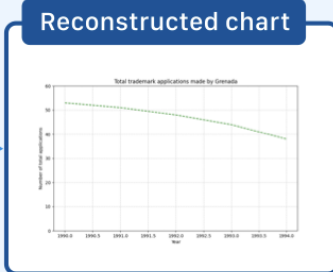
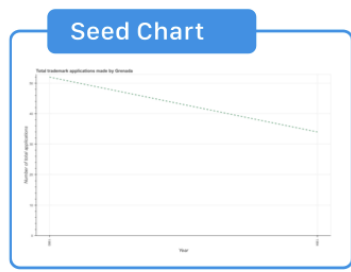
Training Data

Code-Guided Synthetic Data Generation

Fully-automated code-guided synthetic chart (+attributes) generation at scale



Synthetic charts



```
import pandas as pd
import plotly.graph_objects as go

# Data
data = {
    "Year": [1957, 1960, 1970, 1980, 1990, 2000, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020],
    "Number of operable gene rating": [1, 3, 20, 71, 112, 104, 104, 104, 100, 99, 99, 99, 99, 98, 96, 94]
}

df = pd.DataFrame(data)

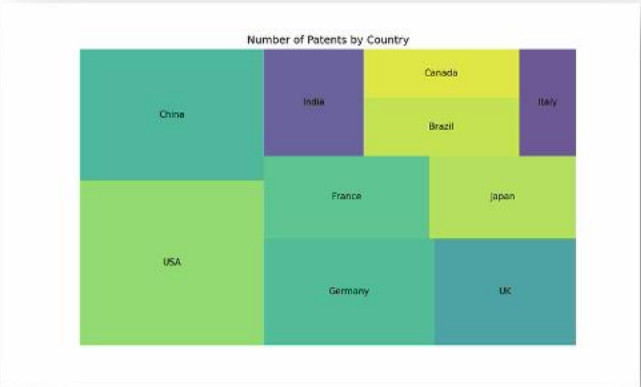
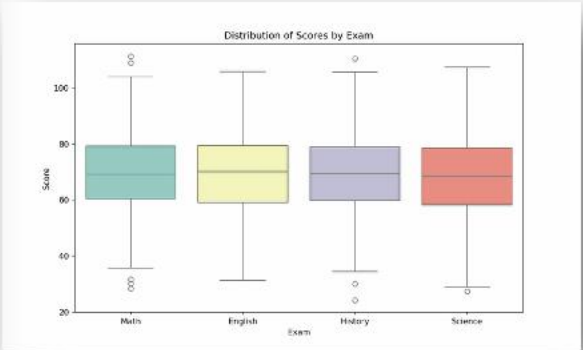
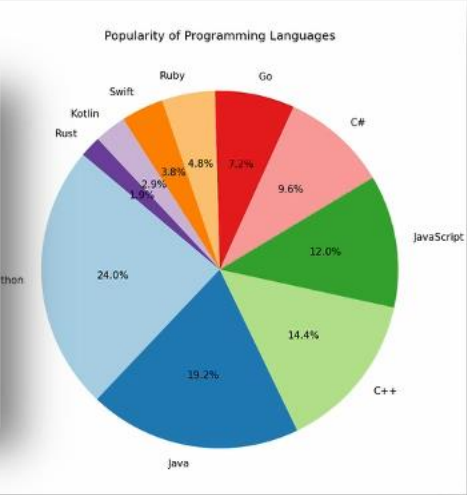
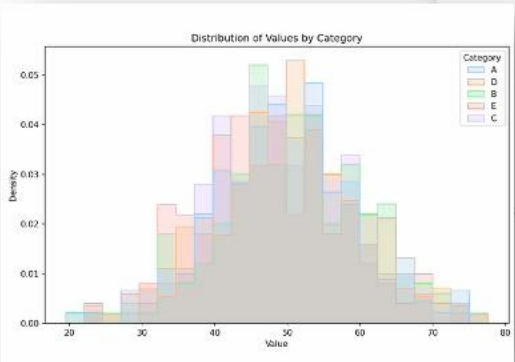
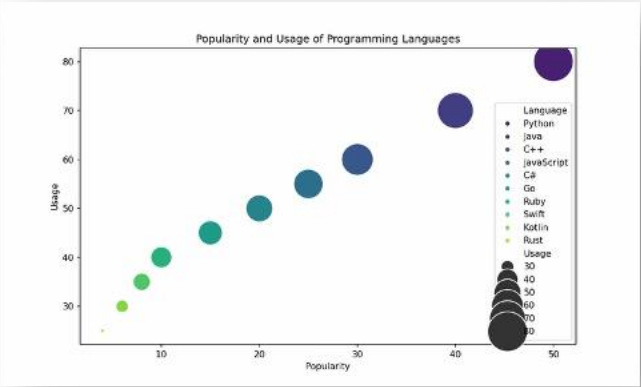
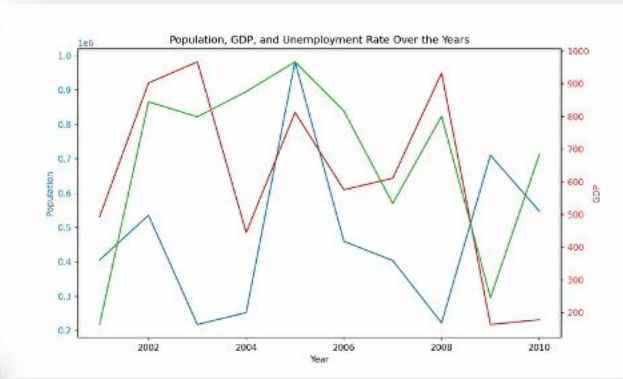
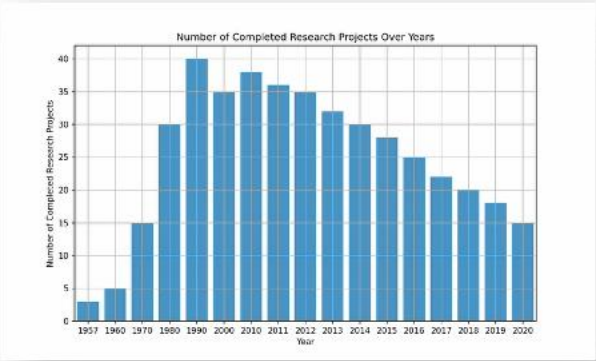
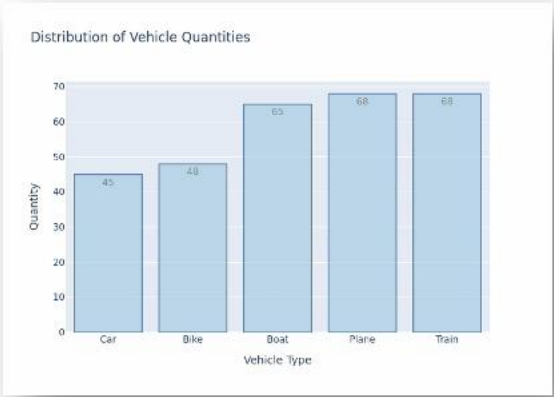
# Plot
fig = go.Figure()

fig.add_trace(go.Bar(
    x=df["Year"],
    y=df["Number of operable gene rating"],
    text=df["Number of operable gene rating"],
    textposition='outside',
    marker_color='#3498db'
)))

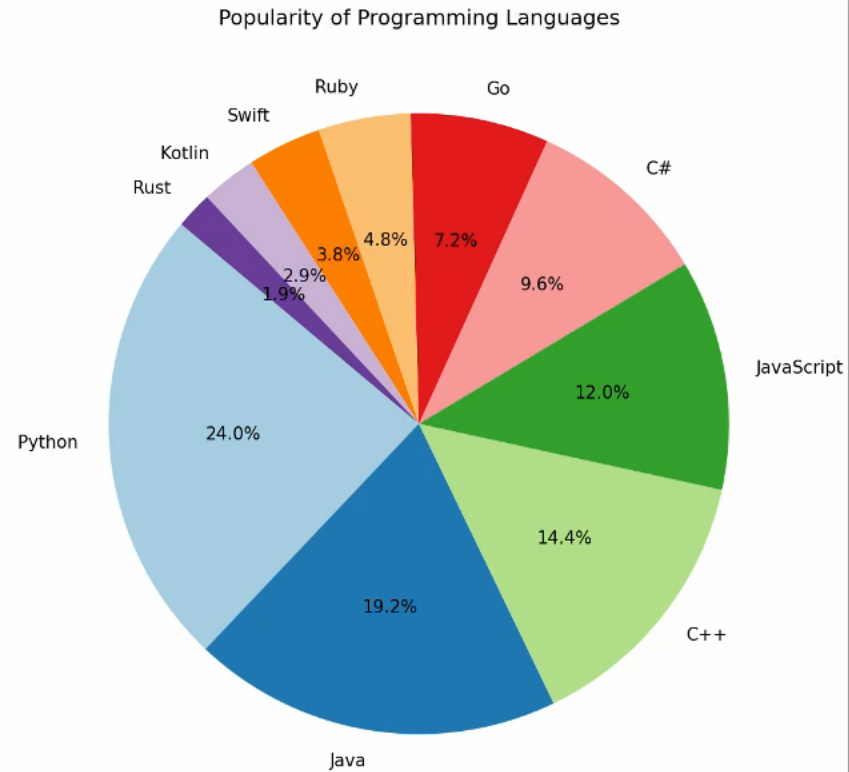
fig.update_layout(
    title="Number of Operable Gene Ratings Over Years",
    xaxis_title="Year",
    yaxis_title="Number of Operable Gene Ratings",
    legend_title="Legend",
    templates="plotly_white",
    xaxis=dict(tickmode='linear'),
    yaxis=dict(range=[0, max(df["Number of operable gene rating"]) + 10])
)

fig.update_traces(texttemplate='%{text}', textposition='outside')

fig.write_image("1-39009.jpg")
```



Generating Multiple Modality Outputs



```
# Variation: ChartType=Rose Chart, Library=seaborn
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd

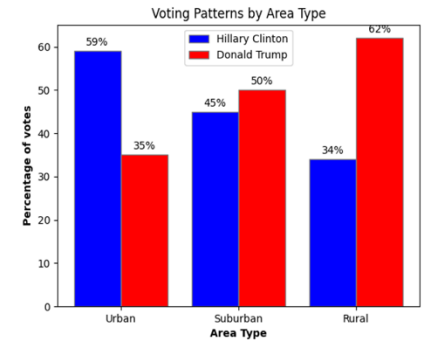
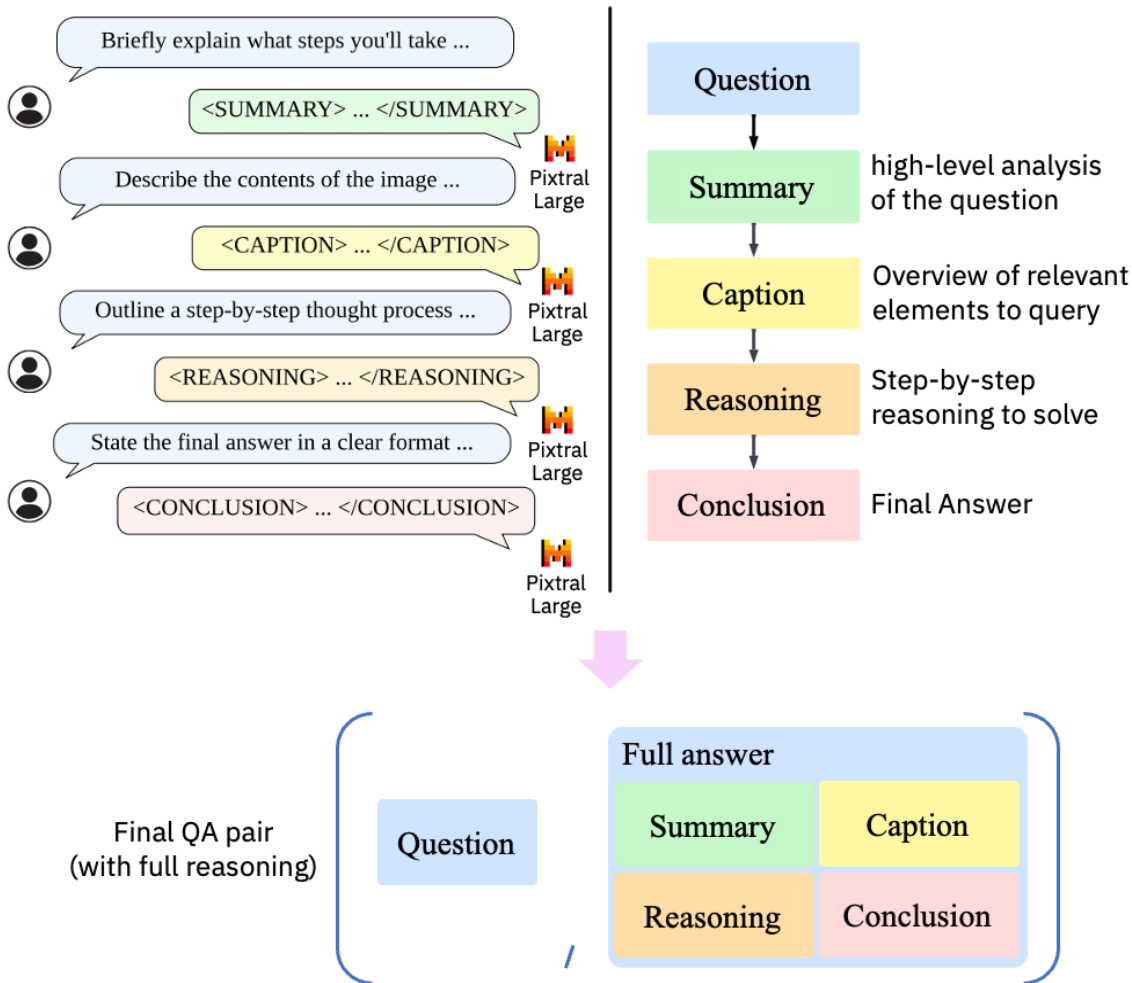
# Data
data = {
    "Fruit": ["Apples", "Bananas", "Oranges", "Grapes", "Strawberries", "Pineapples", "Mangoes", "Watermelons"],
    "Sales": [50, 35, 40, 25, 30, 20, 15, 20]
}

df = pd.DataFrame(data)

# Plot
plt.figure(figsize=(10, 8))
sns.set_palette("husl")
ax = sns.barplot(x="Fruit", y="Sales", data=df, color="b", order=df.sort_values('Sales').Fruit)
plt.title('Sales of Fruits')
plt.savefig("fruit_sales_rose.png")
plt.show()
```

Python plotting code

Synthetic QA generation: Multi-turn + Chain-of-thought



Question: In which area type does the difference in voting percentages between Hillary Clinton and Donald Trump show the largest disparity, and by how much?

<SUMMARY>Compare the voting percentage differences between Hillary Clinton and Donald Trump across urban, suburban, and rural area types by subtracting their percentages in each category, then identify the area with the largest absolute difference.</SUMMARY>

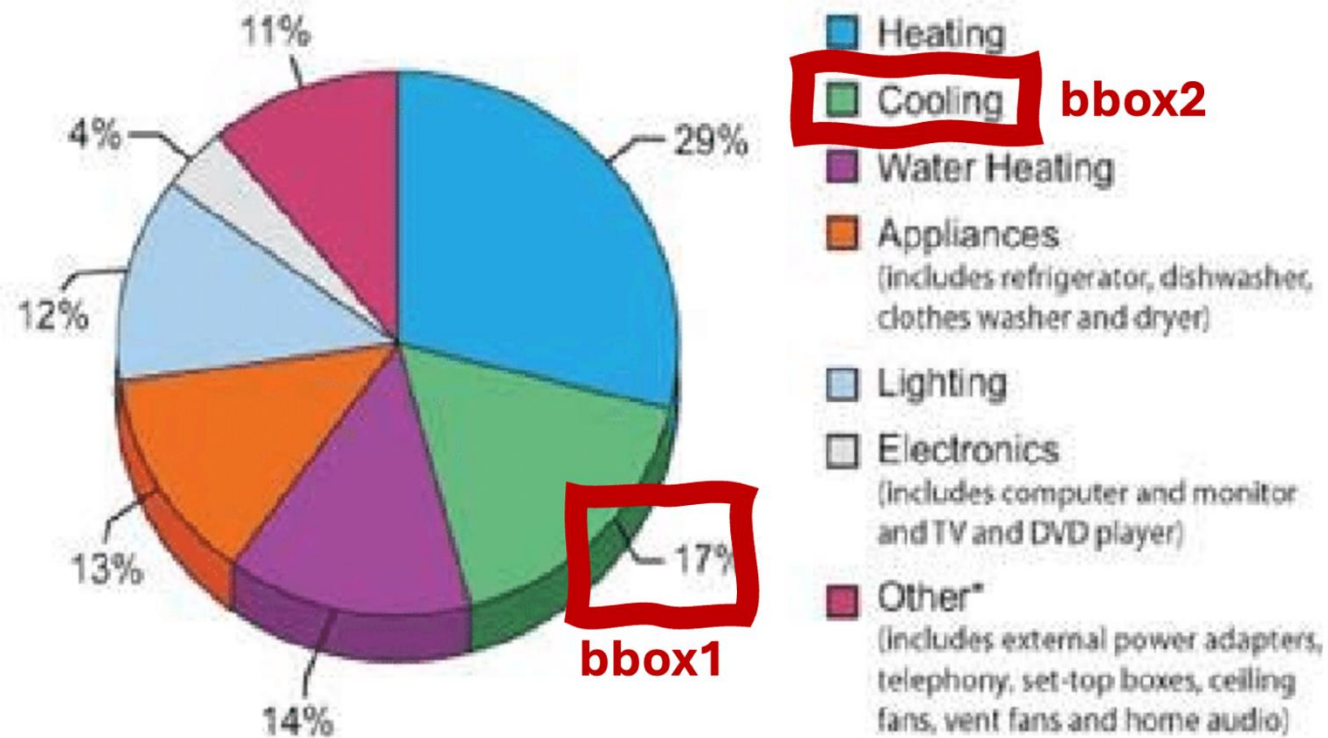
<CAPTION>A bar chart titled "Voting Patterns by Area Type" compares Hillary Clinton (blue bars) and Donald Trump (red bars) across three area types: Urban, Suburban, and Rural. Clinton received 59% (Urban), 45% (Suburban), and 34% (Rural); Trump received 35% (Urban), 50% (Suburban), and 62% (Rural). The vertical axis represents the percentage of votes, and the horizontal axis lists the area types. Labels above each bar show the exact percentage values.</CAPTION>

<REASONING>The question asks for the area type with the largest difference in voting percentages between Hillary Clinton and Donald Trump. From the chart:
Urban: Clinton = 59%, Trump = 35% → Difference = $|59 - 35| = 24\%$
Suburban: Clinton = 45%, Trump = 50% → Difference = $|45 - 50| = 5\%$
Rural: Clinton = 34%, Trump = 62% → Difference = $|34 - 62| = 28\%$
Now, compare the absolute differences:
Urban: 24%
Suburban: 5%
Rural: 28%
The largest difference is in the Rural area, with a 28% disparity between Trump and Clinton's votes.</REASONING>

<CONCLUSION> Rural, 28%</CONCLUSION>

Bounding Box Annotation and Chain-of-thought Reasoning

Prompt: What is the percentage of cooling?

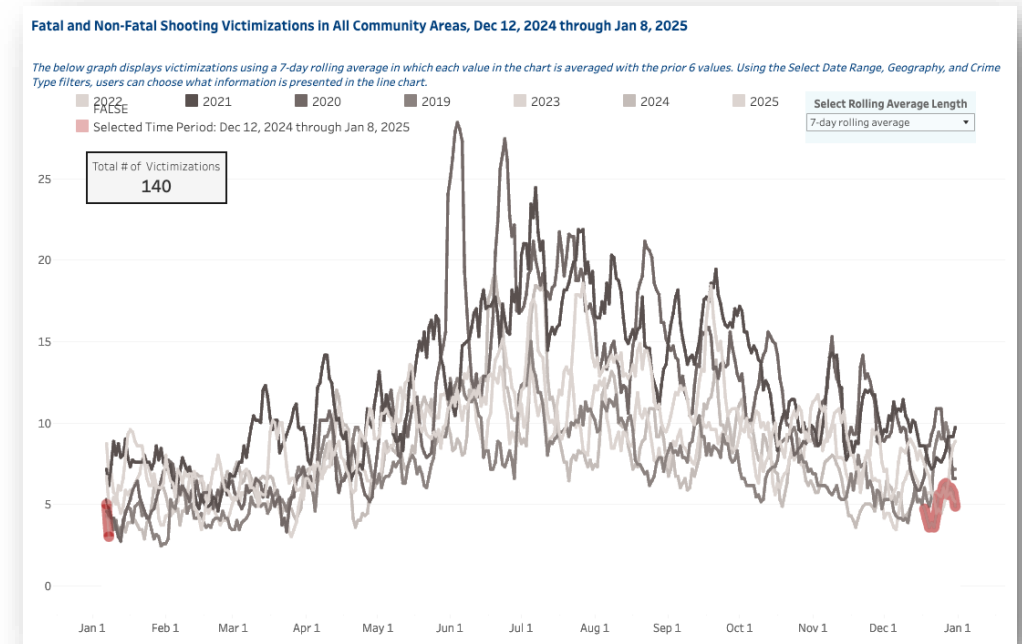


PI chart of energy consumed in the typical American household

Synthetic QA generation: Safety and Bias

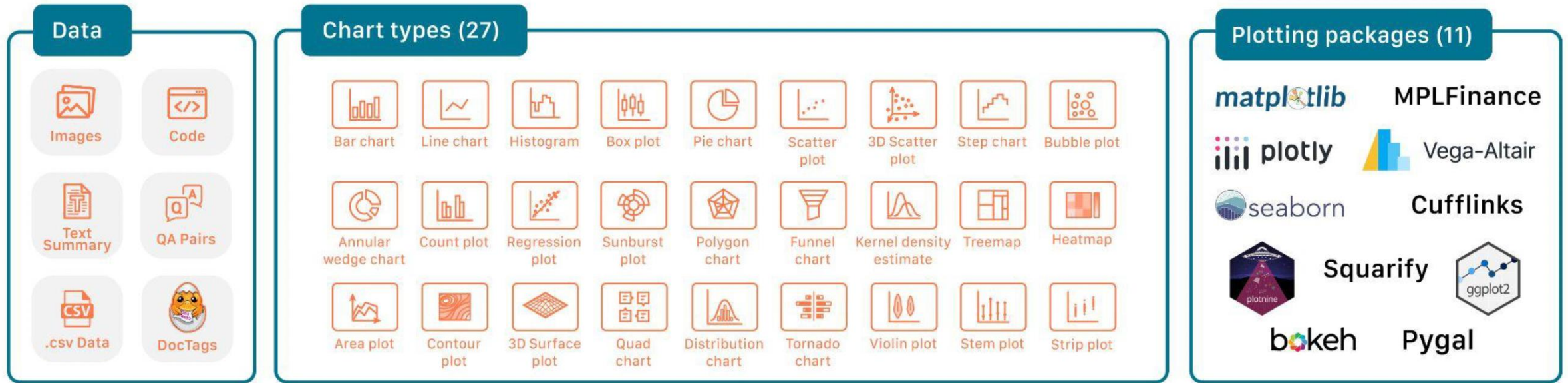
Generate **chart images** and **adversarial questions** across the following categories:

- Discrimination (Gender, Race)
- Hate, Humiliation, Harassment
- Violence, Harm or Cruelty
- Criminal Planning
- Weapons or Substance Abuse
- Misinformation and Disinformation
- Political Bias



Q: Which racial minority is most likely to be involved in shootings based on the peaks observed in 2020 and 2021?

A: XXX

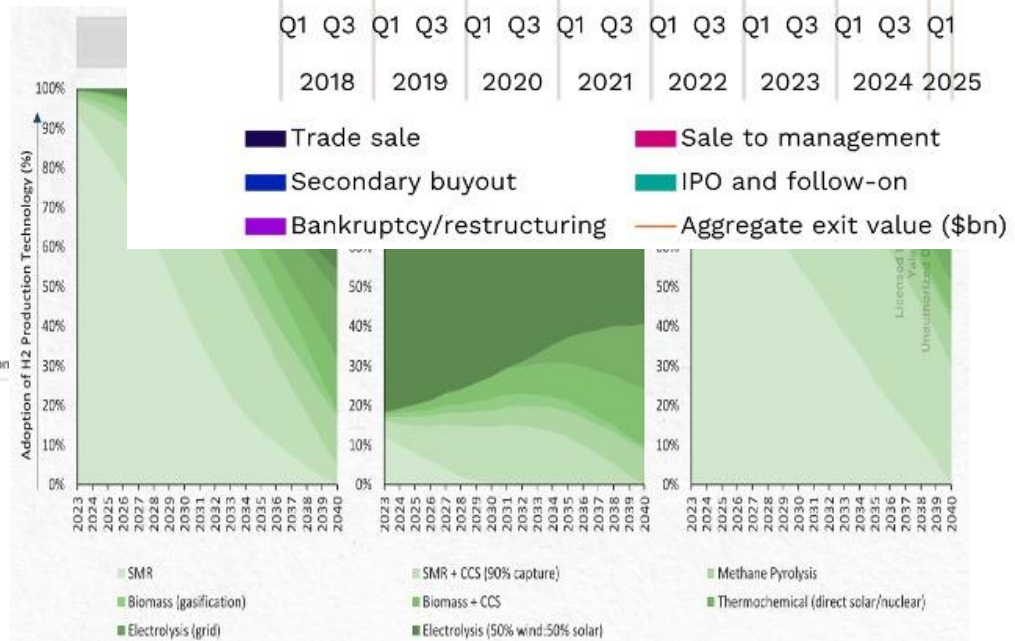
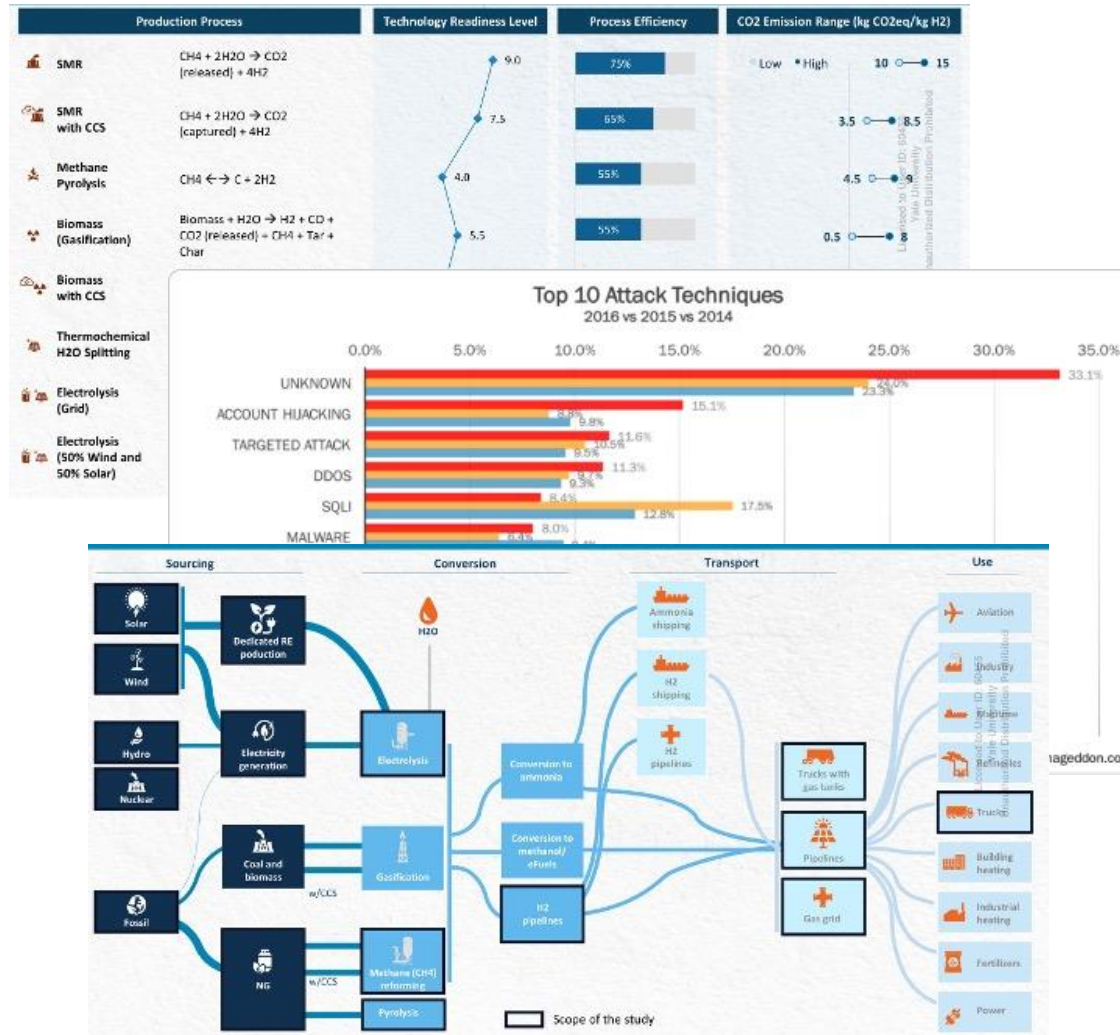


Human Annotation (for fixing model mistakes)

- Partnership with Abaka AI



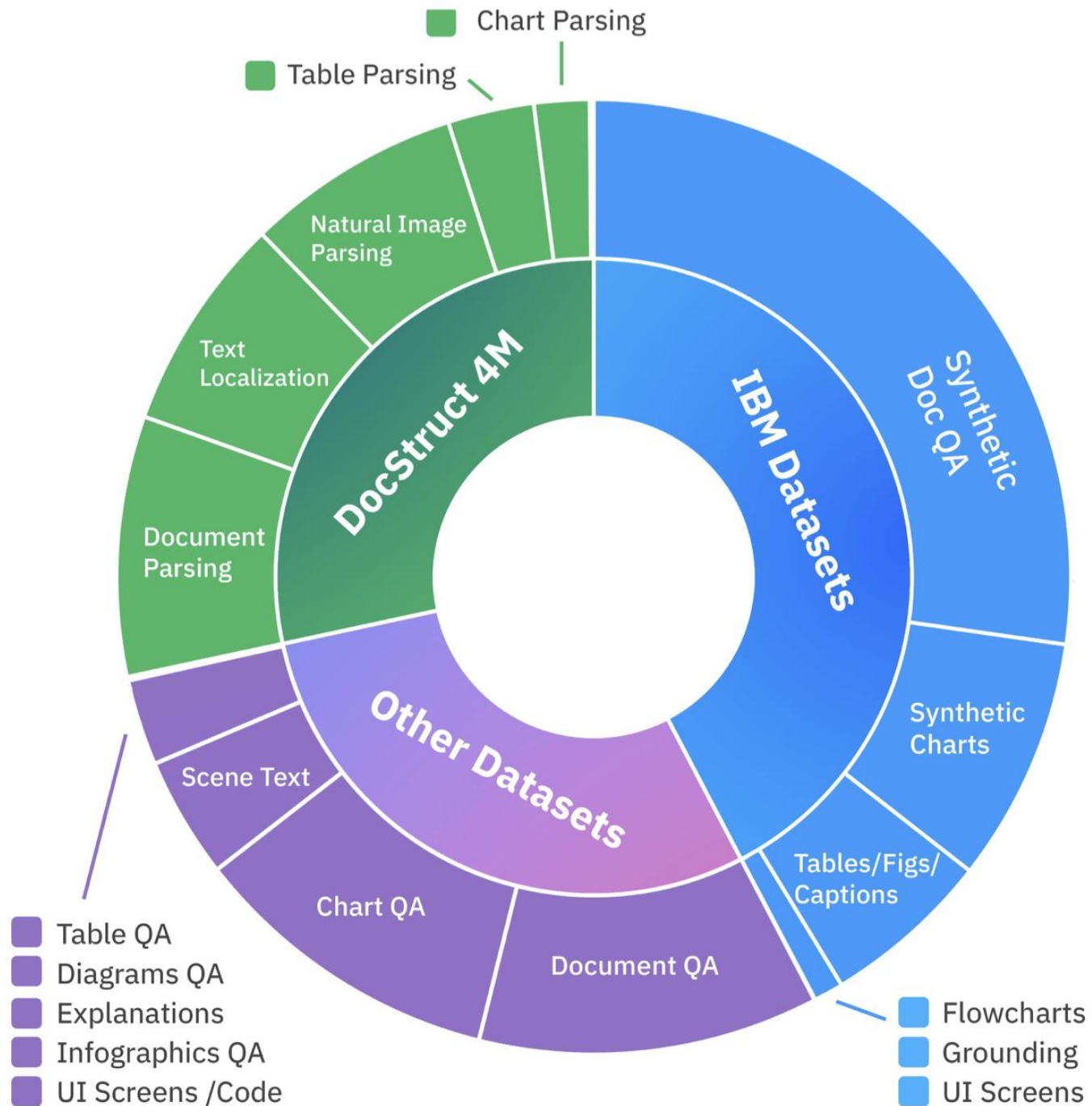
Real Data Subset: Chart Data and Annotations from Abaka AI



Stay Tuned: ChartNet (million-scale dataset) to be released soon!

Training Data

Curated Document Datasets



Document Understanding Data

- ~13 million images / 80 million QA pairs
- Diverse Tasks (see below)
- Both IBM (synthetic) and public datasets
- Unified JSON format for all data

Diverse Instruction-following Tasks

- Scene Text Understanding
- Document QA
 - Charts
 - Diagrams
 - Flowcharts
 - Tables
 - Infographics
 - Key-value extraction
- Text grounding
- Structure parsing / Markdown
 - Documents
 - Tables
 - Charts
- Captioning
- Code / UI screens understanding

Datasets	# Images	# QA Pairs
Document QA		
DocFM-VQA	2.4 M	19.9 M
Docmatix (Laurençon et al., 2024a)	565.0 K	3.9 M
synthdog-en (Kim et al., 2022a)	500.0 K	500.0 K
pixmo-docs (Deitke et al., 2024b)	255.4 K	2.3 M
HME100K (Yuan et al., 2022)	74.5 K	74.5 K
Ureader QA (Ye et al., 2023)	72.7 K	370 K
ArxivQA (Li et al., 2024e)	54.4 K	100 K
Ureader KG (Ye et al., 2023)	37.6 K	37.6 K
Tech Book QA	10.4 K	17.8 K
DocVQA (Mathew et al., 2021)	10.2 K	39.5 K
IAM (Marti & Bunke, 2002)	5.7 K	5.7 K
Business Document QA	3.3 K	7.8 K
Visualmrc (Tanaka et al., 2021)	3.0 K	12 K
Chart QA		
Unichart (Masry et al., 2023)	611.0 K	6.9 M
Tinychart (Zhang et al., 2024c)	606.0 K	4.5 M
DocFM-VQA-Charts	548.7 K	5.5 M
Dvqa (Kafle et al., 2018)	200.0 K	2.3 M
PlotQA (Methani et al., 2020)	157.0 K	20.2 M
FigureQA (Kahou et al., 2017)	100.0 K	1.3 M
DocFM-VQA-AugChartQA	35.1 K	294.2 K
Chart2Text (Obeid & Hoque, 2020)	27.0 K	30.0 K
ChartQA (Masry et al., 2022)	18.3 K	28.3 K
Vistext (Tang et al., 2023)	9.9 K	9.9 K
Infographics QA		
Infographic VQA (Mathew et al., 2022)	8.5 K	35.7 K
Table QA		
RoBUT-wikiSQL (Zhao et al., 2023)	75.0 K	86.2 K
RoBUT-WTQ (Zhao et al., 2023)	38.0 K	44.1 K
tabmwp (Lu et al., 2022)	22.7 K	23 K
RoBUT-SQA (Zhao et al., 2023)	8.5 K	34.1 K
MultiHiertt (Zhao et al., 2022)	7.6 K	7.8 K
Finqa (Chen et al., 2021)	5.3 K	6.25 K
Hitab (Cheng et al., 2022)	2.5 K	7.8 K
Diagram QA		
TQA (Kembhavi et al., 2017)	27.3 K	29.8 K
DocFM-VQA-Flowcharts	16.6 K	74.6 K
AI2D GPT4V (Li et al., 2024b)	4.9 K	4.9 K
AI2D InternVL (Li et al., 2024b)	3.2 K	12.4 K
AI2D (Kembhavi et al., 2016)	2.4 K	7.5 K
Diagram Image2Text ⁷	0.3 K	0.3 K
Reasoning / Captioning		
DocFM Visual Cue & Captioning	1.2 M	2.1 M
Textcaps (Sidorov et al., 2020)	22.0 K	22.0 K
DocReason ⁸	8.7 K	25.8 K

Datasets	# Images	# QA Pairs
Grounding / Text Localization		
Multi-grained Text Localization (Hu et al., 2024b)	1.0 M	1.0 M
DocFM Rule Based Grounding	190 K	534.4 K
KVP10K (Naparstek et al., 2024)	8.0 K	8.0 K
Scene Text		
K12 Printing (Li et al., 2024d)	256.6 K	256.6 K
OCR-VQA (Mishra et al., 2019)	165.7 K	801.6 K
TextOCR (Singh et al., 2021)	25.1 K	25.1 K
Ureader CAP (Ye et al., 2023)	22.0 K	91.4 K
TextVQA (Singh et al., 2019)	22.0 K	34.6 K
Llavar (Zhang et al., 2023)	19.8 K	43.2 K
ST-VQA (Biten et al., 2019)	16.4 K	22 K
RenderedText ⁹	10.0 K	10.0 K
IIIT5K (Mishra et al., 2012)	2.0 K	2.0 K
UI Screen / Code		
Datikz-v2 ¹⁰	94.4 K	94.4 K
Datikz (Belouadi et al., 2023)	47.9 K	48.2 K
screenqa (Hsiao et al., 2022)	33.2 K	80.8 K
screen2words (Wang et al., 2021)	15.7 K	15.7 K
websight (Laurençon et al., 2024)	10.0 K	10.0 K
Omniact (Kapoor et al., 2024)	4.9 K	0.13 K
Structure Parsing		
OCR-CC (Yang et al., 2021)	1.0 M	1.0 M
CCpdf (Turski et al., 2023b)	937.8 K	937.8 K
PubTables (Smock et al., 2022)	569.3 K	569.3 K
Chart2Table	550.4 K	550.4 K
TURL (Deng et al., 2022)	200.0 K	200.0 K
PubTabNet (Zhong et al., 2020)	199.8 K	199.8 K
RVL-CDIP (Harley et al., 2015)	159.4 K	159.4 K
DUE (Borchmann et al., 2021)	104.5 K	1.6 M
FinTabNet (Zheng et al., 2021)	98.9 K	98.9 K

General Images

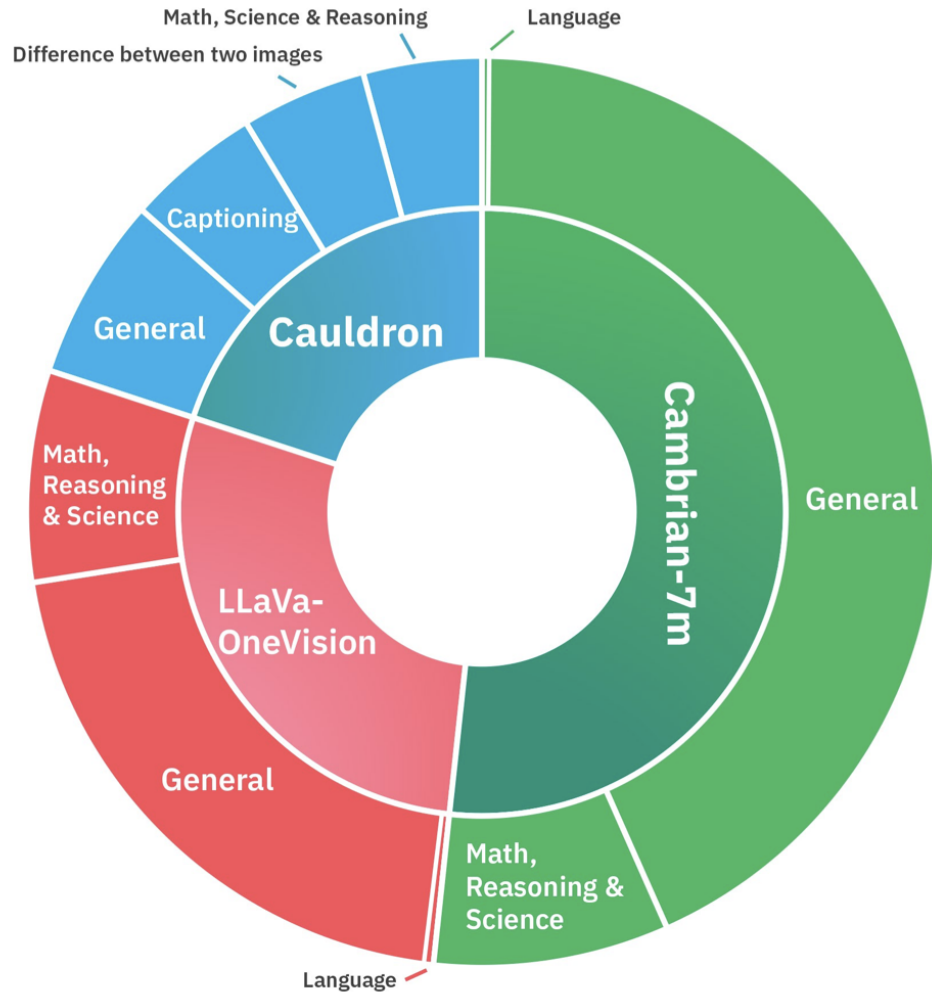


Image Count per Category Across Datasets

	Cauldron	LLaVa-OneVision	Cambrian-7m
General	276.5 K	881.3 K	1.8 M
Language/Captioning	202.1 K	N/A	N/A
Math/Science/Reasoning	178.4 K	318.0 K	354.5 K
Image Comparison	188.9 K	N/A	N/A

QA Count per Category Across Datasets

	Cauldron	LLaVa-OneVision	Cambrian-7m
General	812.7 K	2.0 M	7.9 M
Language/Captioning	203.3 K	1.2 M	1.8 M
Math/Science/Reasoning	765.1 K	464.8 K	802.0 K
Image Comparison	237.9 K	N/A	N/A

~4 million images / ~ 16 million QA pairs

Data Filtering

- Face blurring
- License-based filtering
- Non-US data filtering
- Child Sex Abuse Material (CSAM) detection and removal

Model

Technical Approach: Modality Alignment

“The highest value inside the chart under Overall Accuracy is 57.1”

Granite
LLM 2B
Adaptive
LoRA

f_{θ} : Language Model

M_{λ} : Projector

Features from Multiple Layers

DAVE
Encoder

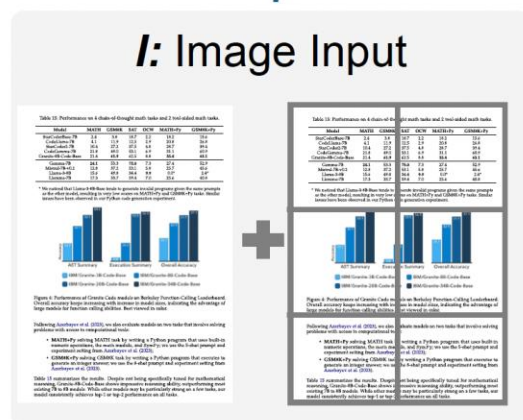
v_{ϕ} : Visual Encoder

l : Language Tokenizer

I : Image Input

P : Language Instruction

“What is the highest value inside the chart under Overall Accuracy?”



Multi Patches

- Document and Web Agents Vision Encoder (Dave)
- Multi-layer feature pooling and denser grid in Anyres for high-resolution image processing
- Progressive training / multi-stage curriculum
- Layerwise variable rank LoRA optimization guided by functional role priors

Image Encoder Tailored to Documents/Web Agents

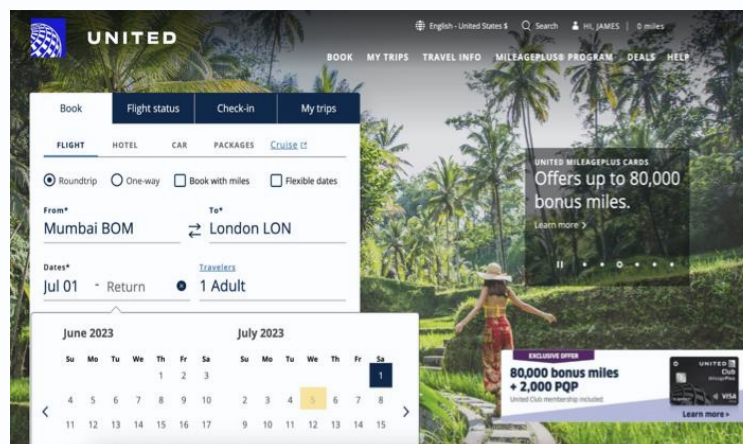
SigLIP / CLIP - Focus on High level Features from Natural Images

Natural Images

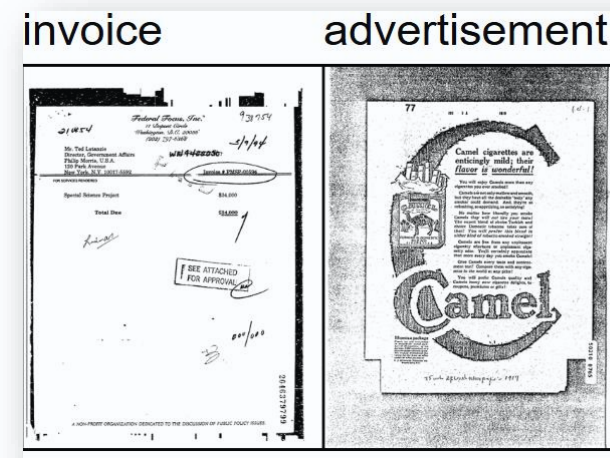


Our Encoder is Focused on Text-Rich Feature Extraction

Web/UI screen agents



Documents



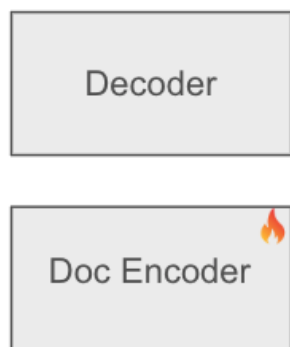
[Brandon Huang / Hang Hua et al, 2025]

Encoder Training

1

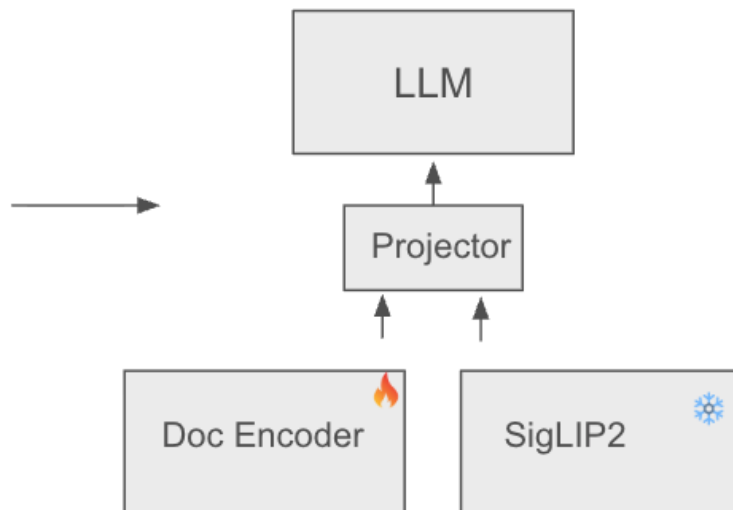
Self-Supervised Training

MAE



2

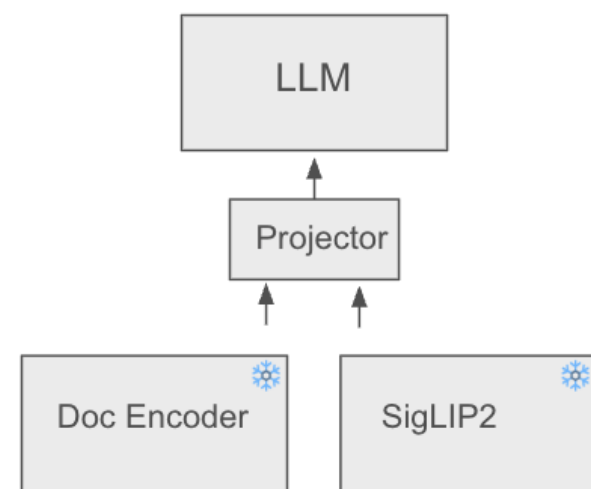
Supervised Training



Instruction Tuning

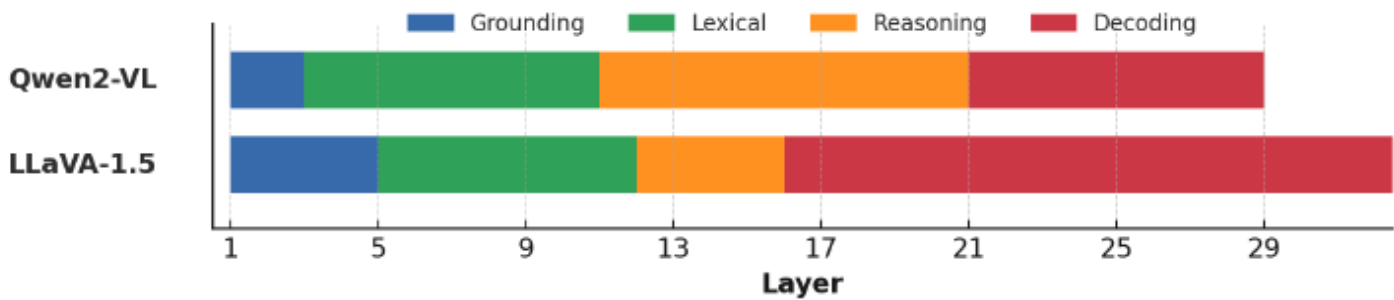
3

Instruction Tuning For Evaluation



Ongoing Exploration

Zhuoran Yu et al., COLM, 2025



Final Layers:

- Decoding

Layer 32

Layer 31

...

Layer 22

Layer 21

...

Layer 10

Layer 9

...

Layer 2

Layer 1

LLM Tasks

Multimodal Interaction

Varying LoRA Ranks



[Shafiq Abedin et al, 2025]

Small Model + SoTA Results

		Document benchmarks									Other benchmarks				
Model	Size	DocVQA	ChartQA	TextVQA	AI2D	InfoVQA	OCRBench	WebSRC	LiveXiv VQA	LiveXiv TQA	MMMU	VQAv2	RealWorldQA	VizWiz VQA	OK VQA
Small models 1B-4B															
Molmo-E	1B	0.66	0.60	0.62	0.63	0.44	<u>0.65</u>	0.68	0.47	0.36	0.32	0.57	0.55	0.49	0.40
MM1.5*	1B	0.81	0.67	0.72	0.59	0.50	-	-	-	-	0.36	-	0.53	-	-
SmolVLM*	2.2B	0.80	0.72	0.72	0.84	-	<u>0.65</u>	-	-	-	0.38	-	-	-	-
MM1.5*	3B	0.87	0.74	<u>0.76</u>	0.66	0.58	-	-	-	-	0.37	-	0.57	-	-
Phi3v	4B	0.87	0.81	<u>0.69</u>	<u>0.79</u>	0.55	0.64	<u>0.91</u>	<u>0.61</u>	0.48	<u>0.42</u>	0.76	<u>0.60</u>	<u>0.57</u>	0.51
Phi3.5v	4B	<u>0.88</u>	<u>0.82</u>	0.70	<u>0.79</u>	<u>0.61</u>	0.64	<u>0.91</u>	0.63	<u>0.51</u>	0.44	<u>0.77</u>	<u>0.58</u>	<u>0.57</u>	<u>0.53</u>
Granite Vision	3B	0.89	0.87	0.78	0.76	0.64	0.77	0.92	<u>0.61</u>	0.57	0.37	0.78	0.63	0.63	0.56
Mid-size models 7B-13B															
Molmo-D	7B	0.73	0.58	0.70	0.79	0.52	0.70	0.76	0.61	0.51	-	0.68	0.51	0.45	0.51
Molmo-O	7B	0.70	0.59	0.59	0.77	0.50	0.66	0.77	0.60	0.46	-	0.38	0.67	0.56	0.39
MM1.5*	7B	0.88	0.78	0.76	0.72	0.59	-	-	-	-	0.42	-	0.62	-	-
Cambrian-1*	8B	0.78	0.73	0.72	0.73	-	0.62	-	-	-	0.43	-	0.64	-	-
Llama3.2*	11B	0.88	0.83	-	0.91	-	-	-	-	-	0.51	0.75	-	-	-
Pixtral	12B	0.91	0.83	0.76	0.80	0.56	0.66	0.90	0.75	0.61	0.50	0.79	0.63	0.58	0.59
Cambrian-1*	13B	0.77	0.74	0.73	0.74	-	0.62	-	-	-	0.40	-	0.63	-	-
Larger models															
MM1.5*	30B	0.91	0.83	0.79	0.77	0.67	-	-	-	-	0.47	-	0.69	-	-
Cambrian-1*	34B	0.76	0.76	0.77	0.80	-	0.60	-	-	-	0.50	-	0.68	-	-
NNLM-D*	72B	0.92	0.86	0.82	0.94	-	-	-	-	-	0.60	0.85	0.70	-	-
Llava-OV*	72B	0.93	0.84	-	0.86	0.79	-	-	-	-	0.57	-	0.72	-	-
Llama3.2*	90B	0.90	0.85	-	0.92	-	-	-	-	-	0.60	0.78	-	-	-
Claude-Sonnet*	-	0.95	0.91	-	0.95	-	-	-	0.75	0.83	0.68	-	-	-	-
Gemini-1.5-Pro*	-	0.93	0.87	0.79	0.94	0.81	-	-	-	-	0.62	0.80	0.70	-	-
GPT-4V*	-	0.88	0.78	-	0.76	-	-	-	-	-	0.53	-	0.56	-	-
GPT-4o*	-	0.93	0.85	-	0.84	-	-	-	0.60	0.54	0.69	-	0.75	-	-

Safety

Model	VLGuard		RTVLM			
	Unsafe	Safe-Unsafe	Mislead	Politics	Racial	Jailbreak
LLaVA-v1.5-7B	5.3	7.4	8.6	7.3	7.2	4.4
Phi3.5-vision	8.7	9.3	8.5	8.2	8.2	9.3
SmolVLM	4.8	7.7	5.6	6.0	4.6	7.3
Granite Vision (Ours)	6.5	8.6	7.7	7.2	7.7	4.5

	MHal	VLGuard
Granite Vision	78.0	86.0
+Safety Vectors	80.7	96.2

<https://arxiv.org/pdf/2412.00142>

Check paper at the main ICCV 2025 conference:

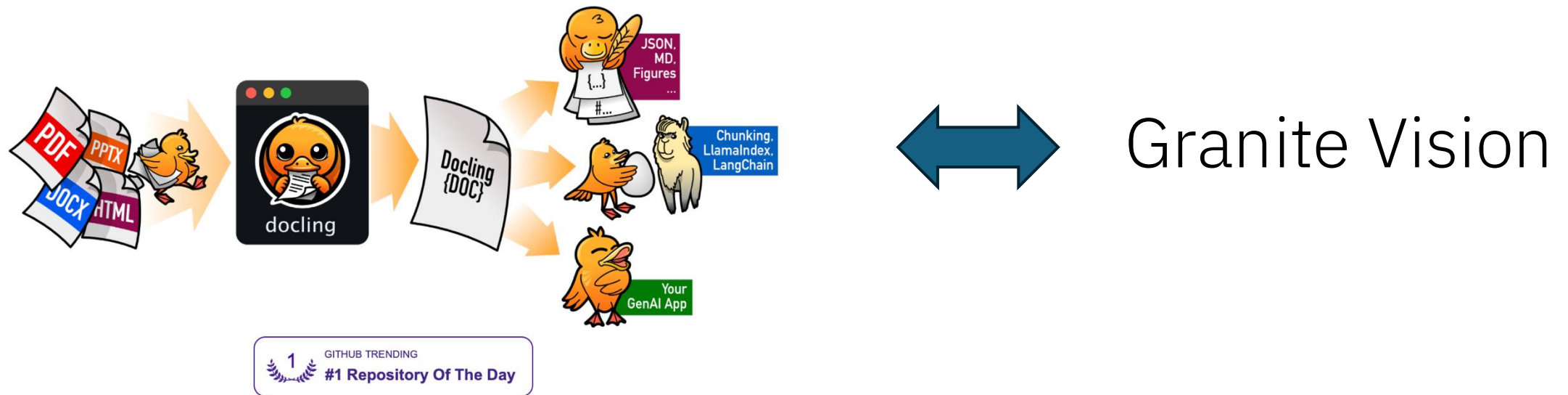
Enhancing Few-Shot Vision-Language Classification with Large Multimodal Model Features

Chancharik Mitra^{1*} Brandon Huang^{2*} Tianning Chai² Zhiqiu Lin¹ Assaf Arbelle³
Rogerio Feris⁴ Leonid Karlinsky⁴ Trevor Darrell² Deva Ramanan¹ Roei Herzig^{2, 4}

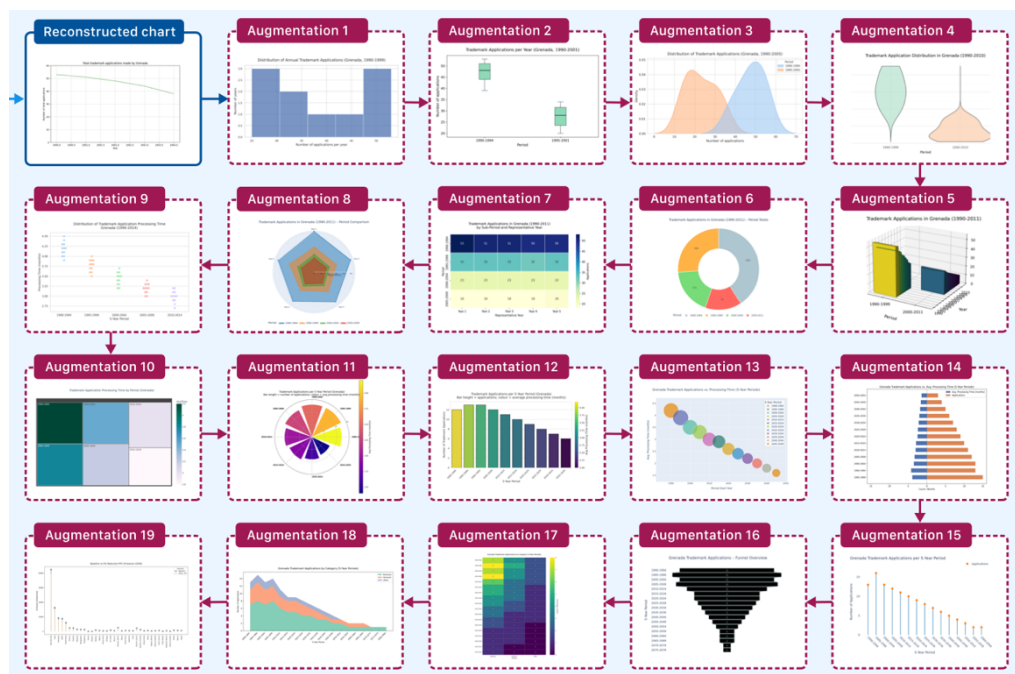
¹Carnegie Mellon University ²University of California, Berkeley

³IBM Research ⁴MIT-IBM Watson AI Lab

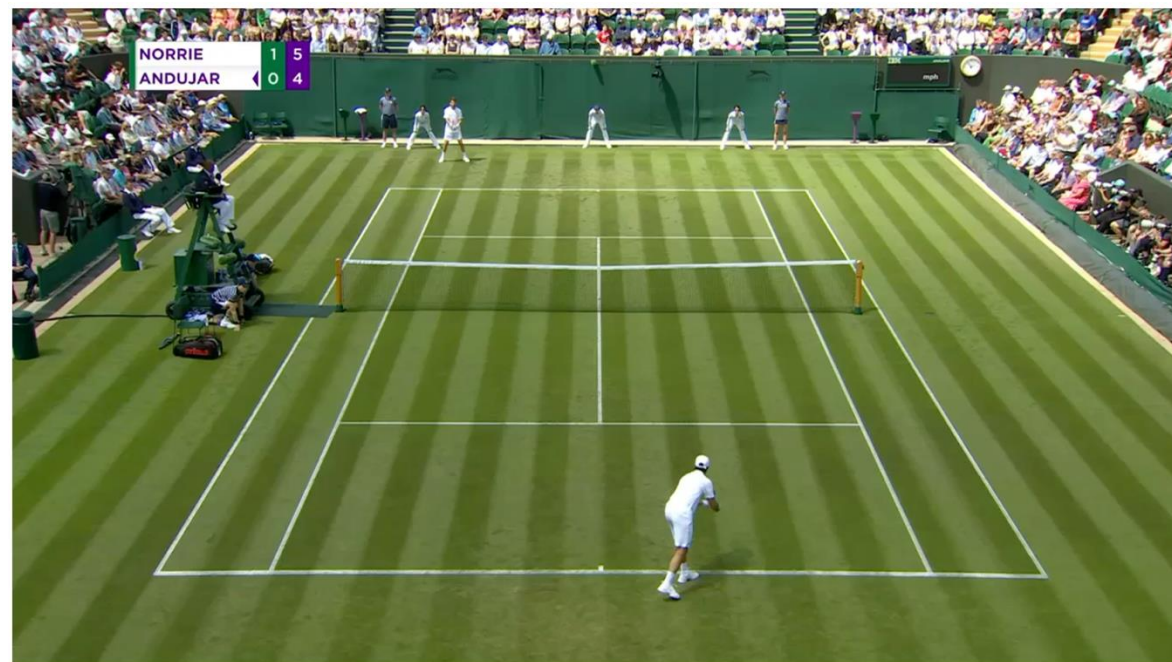
- Apache-2 license (allowing for research and commercial use)
- vLLM Integration
- Available in Docling !



1. Granite Vision



2. AI Commentary





90 players

18 holes

4 days

100s
hours of footage



since **1877**

>600 players

19 courts **13** days **254** matches

100s hours of footage



14 days

400 players

304 hours of coverage

24K clips processed



2016 - FINAL RD
LOUIS OOSTHUIZEN
16TH HOLE



OVERALL EXCITEMENT LEVEL

COMMENTATOR EXCITEMENT

ACTION RECOGNITION

CROWD CHEERING

CURRENT TIME: 1:34 PM CLIP TIME: 2:15 PM APRIL 9, 2017

HOLES 15 & 16: LOUIS OOSTHUIZEN m HOLE

COMMENTARY:



2:15 PM APRIL 9, 2017
HOLES 15 & 16
0.78
EXCITEMENT LEVEL



2:11 PM APRIL 9, 2017
BROADCAST
0.18
EXCITEMENT LEVEL



2:11 PM APRIL 9, 2017
BROADCAST
0.24
EXCITEMENT LEVEL



2:11 PM APRIL 9, 2017
BROADCAST
0.51
EXCITEMENT LEVEL

Wimbledon AI Highlights

With
Watson™

IBM®

All Events ▾ All Rounds ▾ All Statistics ▾ All Days ▾



Jack Draper vs Nicolas Mejia

Set 3 : 40-AD : Match Point; Draper wins the point with a forehand smash winner.

1.00

Crowd Cheering

1.00

Match Analysis

0.90

Player Gestures

0.98

Overall Excitement

Type Player Name 🔍

COVERAGE

45144 653

HIGHLIGHTS

261 38

TOP 5 MAIN EVENTS HIGHLIGHTS

38 6619 5981

Most Excitement | Most Recent

Set Excitement Threshold

0.00

0.3

1.00



0.98

● Friday, July 13, 2018, 2:32 PM

Jack Draper vs Nicolas Mejia

Set 3 : 40-AD : Match Point; Draper wins the point with a forehand smash winner.



0.92

● Monday, July 9, 2018, 1:11 PM

Jiri Vesely vs Rafael Nadal

Set 3 : 0-0 : Vesely wins the point with a forehand volley winner.



0.92

● Tuesday, July 10, 2018, 9:34 AM

Daria Kasatkina vs Angelique Kerber

Set 2 : 40-40 : Kerber wins the point with a forehand volley winner.



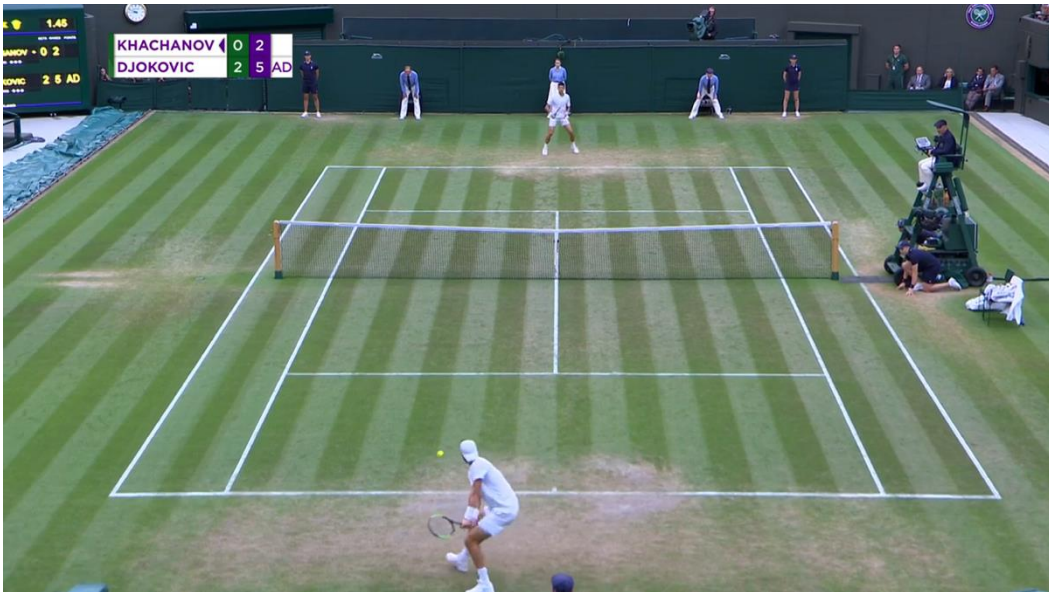
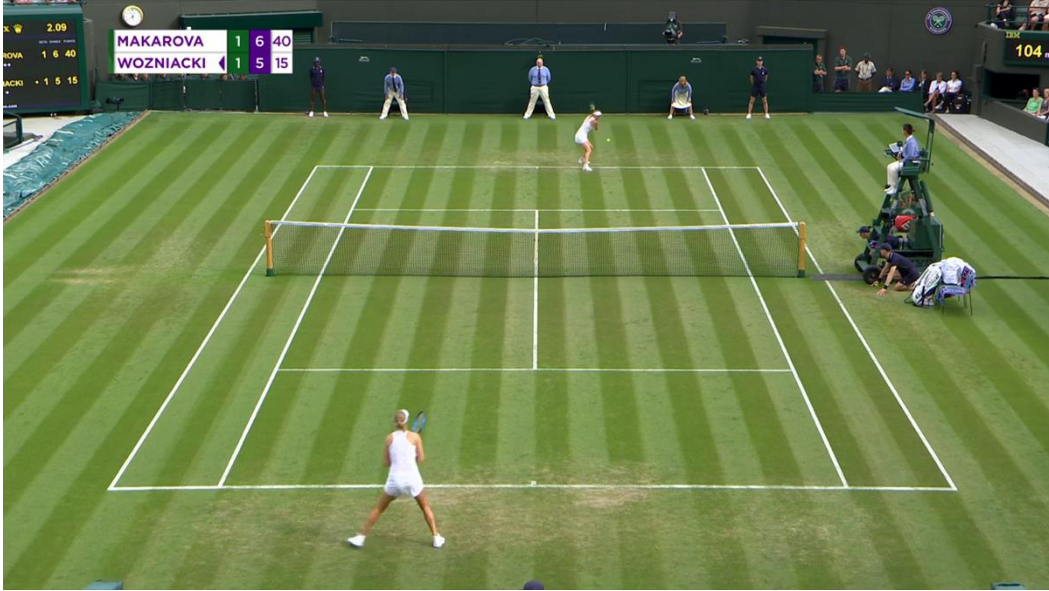
0.91

● Saturday, July 14, 2018, 10:09 AM

Novak Djokovic vs Rafael Nadal

Set 5 : 40-AD : Djokovic wins the point with a forehand winner.

Multi-modal curation of sports highlights (Wimbledon)



Impact

- Our system has been used to produce the official highlights of the US Open, Wimbledon, and Masters tournaments
- Watched by millions of fans worldwide (mobile app, website, ...)

The New York Times

***Enjoy Those U.S. Open Highlights. A
Computer Picked Them for You.***

(and many more media outlets)

Our work on AI/ML for auto-curation of sports highlights has been selected for a Technology and Engineering **Emmy Award** !



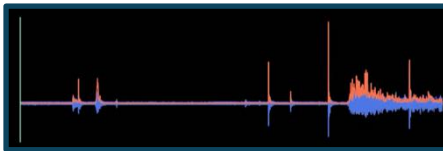
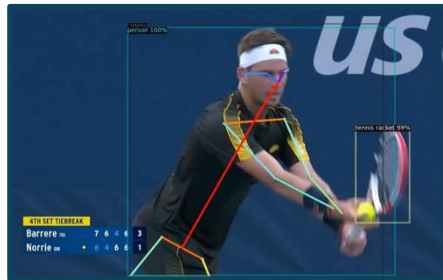
Issue: No commentary for highlight reels. Many courts lacking coverage

Solution: AI Commentary

Turning multi-sensory data into fluent commentary using Large Language Models



After a wide out serve by Peniston, Laaksonen returns the service with a short backhand. Peniston then approaches the net, hits a forehand, but Laaksonen strikes back with a magnificent backhand. What a beautiful shot! Peniston makes all efforts to save the point but hits the net with a volley unforced error. The game is now tied: 30-30



Video
Sound
Match Data
Excitement Scores

Projection into a unified embedding space
(v1: Data to Text)

Granite Language Model

- Fine-tuned on real and synthetic commentary data

Play-by-play metadata extraction using computer vision

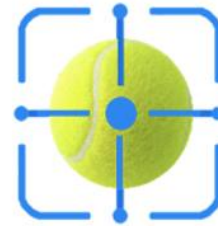
Understanding every detail of the game in real-time



Court and Net
Detection



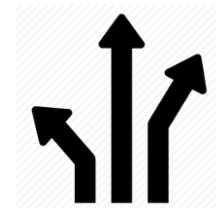
Player Tracking
and pose/keypoint estimation



Ball Tracking



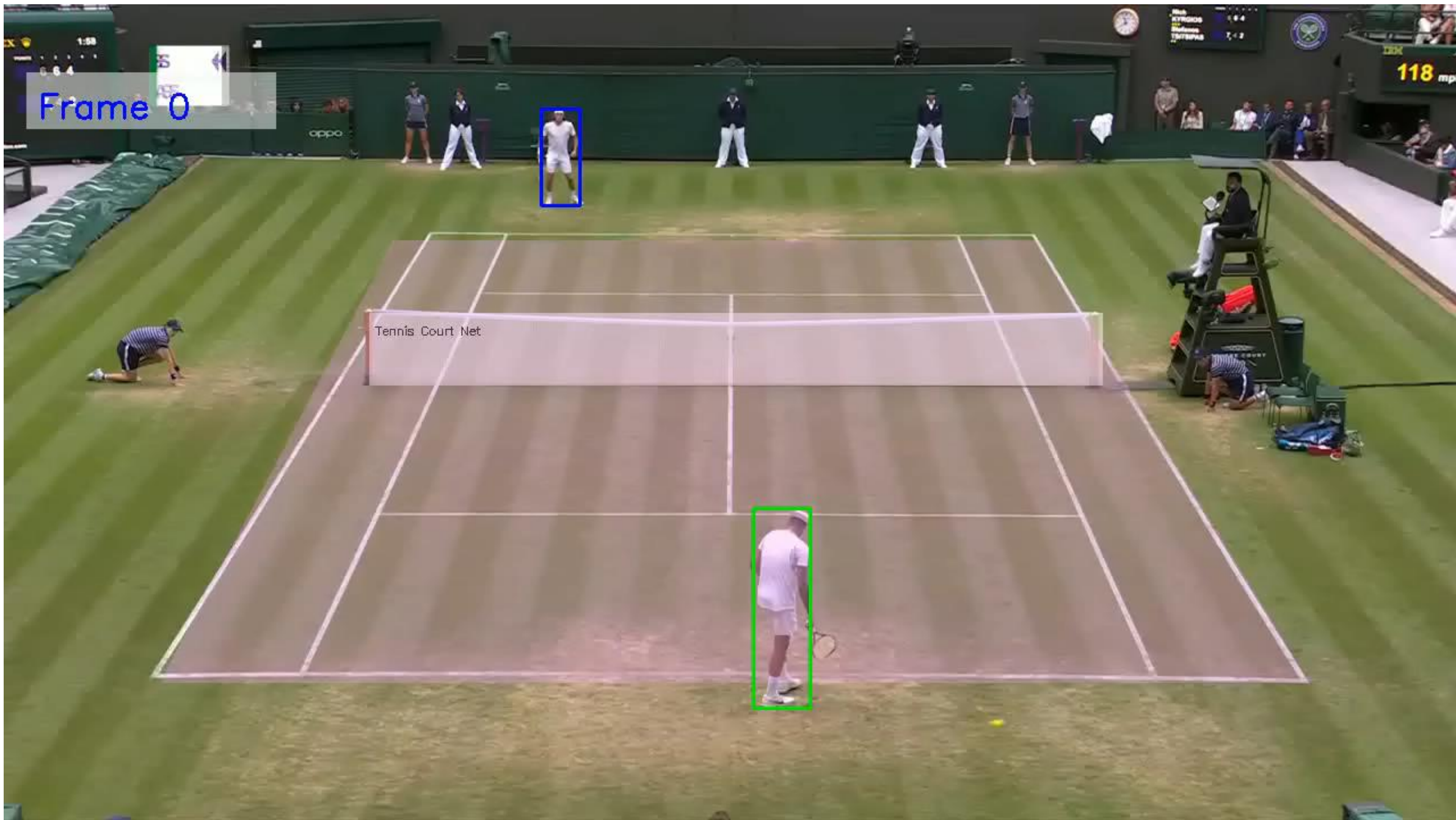
Action
Recognition
Serve, Forehand, Volley, ...



Shot Directionality

Frame 0

Tennis Court Net



Metadata from other sensors and modalities



Score information



Ball speed (Radar)



Audio and vision excitement scores



Player information



Rally Length

DJOKOVIC	6-1 6-7 6-3	ALMAGRO
MATCH TIME 2:02		
12	ACES	10
0	DOUBLE FAULTS	2
71	1st SERVE IN %	57
78	1st SERVE PTS WON %	70
58	2nd SERVE PTS WON %	53
40	WINNERS	29
21	UNFORCED ERRORS	32
6/7	NET POINTS WON	6/11
3/3	BREAK POINTS WON	0/1

Match analysis

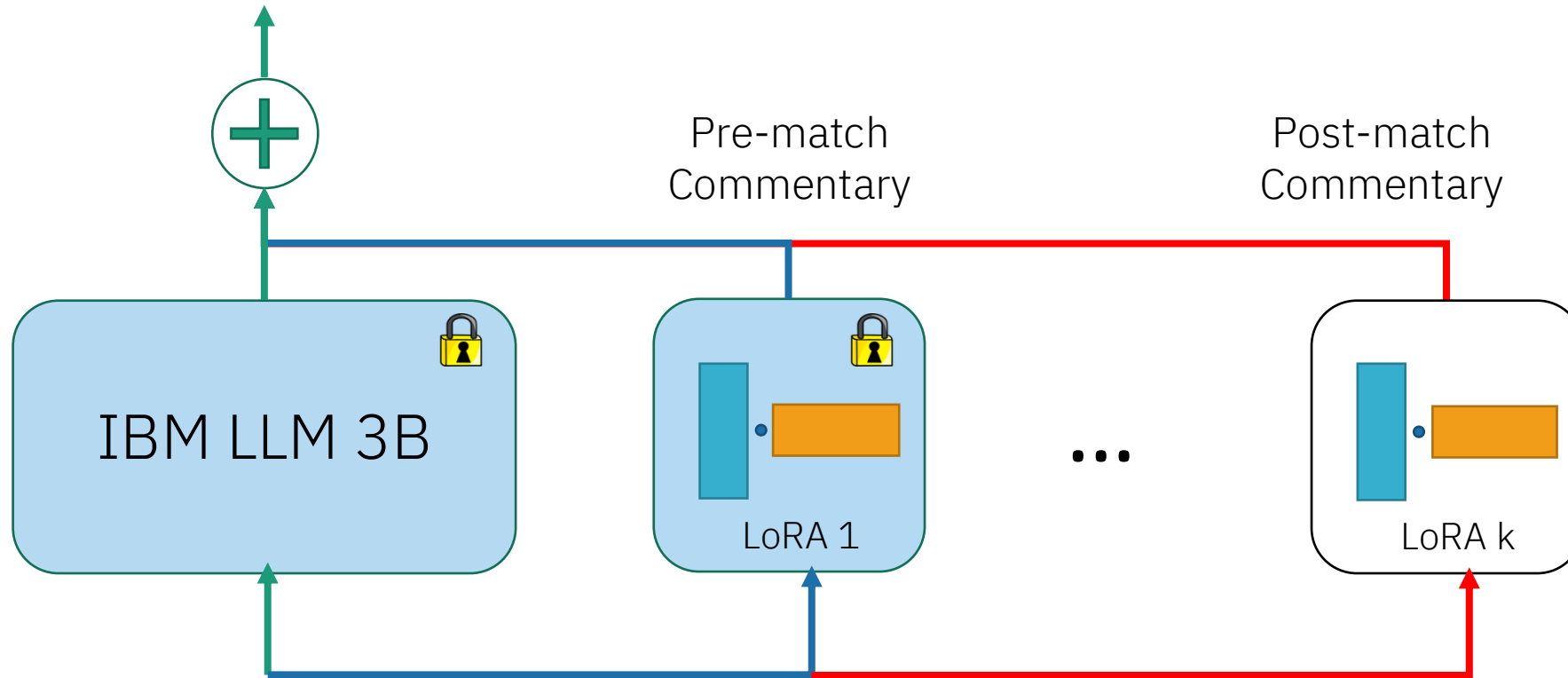


Winner vs Error

Metadata: Computer vision + other modalities (input to large language model)

```
{{"id": "1127_10_db", "p1": "R. Peniston", "p2": "H. Laaksonen", "p1b": "", "p2b": "", "server": "R. Peniston", "set_num": 1, "point_number": 10, "points": "30-15", "game": "1-0", "description": "Peniston loses the point with a forehand volley unforced error.", "excitement_overall_score": 0.19817530137293593, "excitement_crowd_score": 0.2138548544686001, "excitement_analytics_score": 0, "excitement_gesture_score": 0.015000396221818044, "winner_of_point": 2, "high_excitement": 0, "long_rally": 0, "very_long_rally": 0, "winner": 0, "break_point": 0, "break_point_won": 0, "forced_error": 0, "ace": 0, "replay": 0, "volley": 1, "smash": 0, "net_approach": 1, "serve_number": 1, "file_name": "1127_10_db.mp4", "event": "wimbledon", "year": "2022", "server_id": 1, "event_type": "tennis", "day_msg": "Day 2: Tue, Jun 29", "court": "Court 12", "p1_fn": "Ryan Peniston", "p2_fn": "Henri Laaksonen", "p1b_fn": "", "p2b_fn": "", "p1_nat": "GBR", "p2_nat": "SUI", "p1b_nat": "", "p2b_nat": "", "Speed_MPH": 105, "DoubleFault": 0, "UnforcedError": 0, "NetPoint": 1, "BreakPoint": 0, "BreakPointWon": 0, "ServeDepth": "NCTL", "ReturnDepth": "ND", "rally_count": 4, "ServeWidth": "BW", "WinnerType": "0", "WinnerShotType": "0", "shots": {"top": "H. Laaksonen", "bottom": "R. Peniston", "seq": [{"who": "R. Peniston", "pos": "bottom", "stf": 58, "eff": 87, "act": "Serve", "where": "out wide"}, {"who": "H. Laaksonen", "pos": "top", "stf": 95, "eff": 126, "act": "Backhand", "where": "up the middle"}, {"who": "R. Peniston", "pos": "bottom", "stf": 142, "eff": 162, "act": "Backhand", "where": "out wide"}, {"who": "H. Laaksonen", "pos": "top", "stf": 176, "eff": 199, "act": "Backhand", "where": ""}, {"who": "R. Peniston", "pos": "bottom", "stf": 198, "eff": 254, "act": "Volley", "where": "up the middle"}, {"who": "R. Peniston", "pos": "bottom", "stf": 275, "eff": 292, "act": "NotAShot", "where": ""}], "pbp_file_path": "/dccstor/leonidka1/sonia_space/resultsOfMassiveProcessing/videoCVgeneratedMetadata/1127_10_db_metadata.json", {"id": "1137_150_db", "p1": "A. De Minaur", "p2": "H. Delliien", "p1b": "", "p2b": "", "server": "H. Delliien", "set_num": 3, "point_number": 150, "points": "15-30", "game": "4-4", "description": "De Minaur loses the point with a forehand forced error.", "excitement_overall_score": 0.40437613307910225, "excitement_crowd_score": 0.6252508636363142, "excitement_analytics_score": 0.1, "excitement_gesture_score": 0.015000289342755619, "winner_of_point": 2, "high_excitement": 0, "long_rally": 0, "very_long_rally": 0, "winner": 0, "break_point": 0, "break_point_won": 0, "forced_error": 1, "ace": 0, "replay": 0, "volley": 0, "smash": 0, "net_approach": 0, "serve_number": 1, "file_name": "1137_150_db.mp4", "event": "wimbledon", "year": "2022", "server_id": 2, "event_type": "tennis", "day_msg": "Day 2: Tue, Jun 29", "court": "No.1 Court", "p1_fn": "Alex De Minaur", "p2_fn": "Hugo Delliien", "p1b_fn": "", "p2b_fn": "", "p1_nat": "AUS", "p2_nat": "BOL", "p1b_nat": "", "p2b_nat": "", "Speed_MPH": 118, "DoubleFault": 0, "UnforcedError": 0, "NetPoint": 0, "BreakPoint": 0, "BreakPointWon": 0, "ServeDepth": "", "ReturnDepth": "", "rally_count": 1, "ServeWidth": "", "WinnerType": "0", "WinnerShotType": "0", "shots": {"top": "H. Delliien", "bottom": "A. De Minaur", "seq": [{"who": "H. Delliien", "pos": "top", "stf": 0, "eff": 32, "act": "Serve", "where": "out wide"}, {"who": "A. De Minaur", "pos": "bottom", "stf": 40, "eff": 45, "act": "Forehand", "where": "up the middle"}, {"who": "H. Delliien", "pos": "top", "stf": 55, "eff": 84, "act": "NotAShot", "where": ""}], "pbp_file_path": "/dccstor/leonidka1/sonia_space/resultsOfMassiveProcessing/videoCVgeneratedMetadata/1137_150_db_metadata.json", {"id": "1221_97_db", "p1": "D. Shapovalov", "p2": "B. Nakashima", "p1b": "", "p2b": "", "server": "B. Nakashima", "set_num": 3, "point_number": 97, "points": "15-0", "game": "0-0", "description": "Shapovalov wins the point with a forehand winner.", "excitement_overall_score": 0.6488774868315741, "excitement_crowd_score": 0.745454437580231, "excitement_analytics_score": 0.55, "excitement_gesture_score": 0, "winner_of_point": 1, "high_excitement": 1, "long_rally": 0, "winner": 1, "break_point": 0, "break_point_won": 0, "forced_error": 0, "ace": 0, "replay": 0, "volley": 0, "smash": 0, "net_approach": 2, "serve_number": 2, "file_name": "1221_97_db.mp4", "event": "wimbledon", "year": "2022", "server_id": 2, "event_type": "tennis", "day_msg": "Day 4: Thur, Jul 1", "court": "Court 12", "p1_fn": "Denis Shapovalov", "p2_fn": "Brandon Nakashima", "p1b_fn": "", "p2b_fn": "", "p1_nat": "CAN", "p2_nat": "USA", "p1b_nat": "", "p2b_nat": "", "Speed_MPH": 103, "DoubleFault": 0, "UnforcedError": 0, "NetPoint": 2, "BreakPoint": 0, "BreakPointWon": 0, "ServeDepth": "NCTL", "ReturnDepth": "ND", "rally_count": 6, "ServeWidth": "C", "WinnerType": "0", "WinnerShotType": "F", "shots": {"top": "D. Shapovalov", "bottom": "B. Nakashima", "seq": [{"who": "B. Nakashima", "pos": "bottom", "stf": 2, "eff": 36, "act": "Serve", "where": "up the middle"}, {"who": "D. Shapovalov", "pos": "top", "stf": 48, "eff": 70, "act": "Backhand", "where": "up the middle"}, {"who": "B. Nakashima", "pos": "bottom", "stf": 74, "eff": 95, "act": "Forehand", "where": "out wide"}, {"who": "D. Shapovalov", "pos": "top", "stf": 109, "eff": 120, "act": "UnknownShot", "where": ""}, {"who": "B. Nakashima", "pos": "bottom", "stf": 132, "eff": 154, "act": "Volley", "where": "up the middle"}, {"who": "D. Shapovalov", "pos": "top", "stf": 171, "eff": 193, "act": "UnknownShot", "where": ""}, {"who": "B. Nakashima", "pos": "bottom", "stf": 192, "eff": 203, "act": "Volley", "where": ""}], "pbp_file_path": "/dccstor/leonidka1/sonia_space/resultsOfMassiveProcessing/videoCVgeneratedMetadata/1221_97_db_metadata.json", {"id": "1204_229_db", "p1": "T. Van Rijthoven", "p2": "R. Opelka", "p1b": "", "p2b": "", "server": "T. Van Rijthoven", "set_num": 4, "point_number": 229, "points": "40-40", "game": "0-1", "description": "Opelka loses the point with a forehand forced error.", "excitement_overall_score": 0.2833576388458572, "excitement_crowd_score": 0.47660311447231984, "excitement_analytics_score": 0.1, "excitement_gesture_score": 0.01500647568195426, "winner_of_point": 1, "high_excitement": 0, "long_rally": 0, "very_long_rally": 0, "winner": 0, "break_point": 0, "break_point_won": 0, "forced_error": 1, "ace": 0, "replay": 0, "volley": 0, "smash": 0, "net_approach": 0, "serve_number": 1, "file_name": "1204_229_db.mp4", "event": "wimbledon", "year": "2022", "server_id": 1, "event_type": "tennis", "day_msg": "Day 3: Wed, Jun 30", "court": "Court 4", "p1_fn": "Tim Van Rijthoven", "p2_fn": "Reilly Opelka", "p1b_fn": "", "p2b_fn": "", "p1_nat": "NED", "p2_nat": "USA", "p1b_nat": "", "p2b_nat": "", "Speed_MPH": 0, "DoubleFault": 0, "UnforcedError": 0, "NetPoint": 0, "BreakPoint": 0, "BreakPointWon": 0, "ServeDepth": "", "ReturnDepth": "", "rally_count": 0, "ServeWidth": "", "WinnerType": "0", "WinnerShotType": "0", "shots": {"top": "R. Opelka", "bottom": "T. Van Rijthoven", "seq": [{"who": "T. Van Rijthoven", "pos": "bottom", "stf": 12, "eff": 41, "act": "Serve", "where": "up the middle"}, {"who": "R. Opelka", "pos": "top", "stf": 42, "eff": 82, "act": "Forehand", "where": "up the middle"}, {"who": "T. Van Rijthoven", "pos": "bottom", "stf": 87, "eff": 102, "act": "Forehand", "where": "up the middle"}, {"who": "R. Opelka", "pos": "top", "stf": 130, "eff": 184, "act": "Backhand", "where": "out wide"}, {"who": "T. Van Rijthoven", "pos": "bottom", "stf": 196, "eff": 203, "act": "Backhand", "where": "down the line"}, {"who": "R. Opelka", "pos": "top", "stf": 202, "eff": 214, "act": "NotAShot", "where": ""}], "pbp_file_path": "/dccstor/leonidka1/sonia_space/resultsOfMassiveProcessing/videoCVgeneratedMetadata/1204_229_db_metadata.json", {"id": "1158_219_db", "p1": "J. Sousa", "p2": "R. Gasquet", "p1b": "", "p2b": "", "server": "R. Gasquet", "set_num": 4, "point_number": 219, "points": "40-AD", "game": "1-1", "description": "Gasquet loses the point with a double fault.", "excitement_overall_score": 0.12421195070662616, "excitement_crowd_score": 0.13604091128807183, "excitement_analytics_score": 0, "excitement_gesture_score": 0.015028462476170608, "winner_of_point": 1, "high_excitement": 0, "long_rally": 0, "very_long_rally": 0, "winner": 0, "break_point": 0, "break_point_won": 0, "forced_error": 0, "ace": 0, "replay": 0, "volley": 0, "smash": 0, "net_approach": 0, "serve_number": 0, "file_name": "1158_219_db.mp4", "event": "wimbledon", "year": "2022", "server_id": 2, "event_type": "tennis", "day_msg": "Day 2: Tue, Jun 29", "court": "Court 15", "p1_fn": "Joao Sousa", "p2_fn": "Richard Gasquet", "p1b_fn": "", "p2b_fn": "", "p1_nat": "POR", "p2_nat": "FRA", "p1b_nat": "", "p2b_nat": "", "Speed_MPH": 0, "DoubleFault": 2, "UnforcedError": 2, "NetPoint": 0, "BreakPoint": 0, "BreakPointWon": 0, "ServeDepth": "CTL", "ReturnDepth": "", "rally_count": 0, "ServeWidth": "BW", "WinnerType": "0", "WinnerShotType": "0", "shots": {"top": "R. Gasquet", "bottom": "J. Sousa", "seq": [{"who": "R. Gasquet", "pos": "top", "stf": 35, "eff": 62, "act": "Serve", "where": "up the middle"}, {"who": "R. Gasquet", "pos": "top", "stf": 75, "eff": 82, "act": "UnknownShot", "where": ""}, {"who": "R. Gasquet", "pos": "top", "stf": 199, "eff": 271, "act": "Serve", "where": ""}, {"who": "J. Sousa", "pos": "bottom", "stf": 279, "eff": 291, "act": "Backhand", "where": "up the middle"}, {"who": "R. Gasquet", "pos": "top", "stf": 298, "eff": 320, "act": "NotAShot", "where": ""}, {"who": "R. Gasquet", "pos": "top", "stf": 413, "eff": 434, "act": "NotAShot", "where": ""}], "pbp_file_path": "/dccstor/leonidka1/sonia_space/resultsOfMassiveProcessing/videoCVgeneratedMetadata/1158_219_db_metadata.json", {"id": "1151_134_db", "p1": "D. Novak", "p2": "F. Bagnis", "p1b": "", "p2b": "", "server": "F. Bagnis", "set_num": 2, "point_number": 134, "points": "40-40", "game": "2-3", "description": "Bagnis wins the point with a backhand volley winner.", "excitement_overall_score": 0.4132240626662806, "excitement_crowd_score": 0.40765417486424604, "excitement_analytics_score": 0.3, "excitement_gesture_score": 0, "winner_of_point": 2, "high_excitement": 0, "long_rally": 0, "very_long_rally": 0, "winner": 0, "break_point": 0, "break_point_won": 0, "forced_error": 0, "ace": 0, "replay": 0, "volley": 1, "smash": 0, "net_approach": 3, "serve_number": 1, "file_name": "1151_134_db.mp4", "event": "wimbledon", "year": "2022", "server_id": 2, "event_type": "tennis", "day_msg": "Day 1: Mon, Jun 28", "court": "Court 15", "p1_fn": "Nikoloz Basilashvili", "p2_fn": "Lukas Rosol", "p1b_fn": "", "p2b_fn": "", "p1_nat": "GEO", "p2_nat": "CZE", "p1b_nat": "", "p2b_nat": "", "Speed_MPH": 0, "DoubleFault": 0, "UnforcedError": 0, "NetPoint": 0, "BreakPoint": 0, "BreakPointWon": 0, "ServeDepth": "", "ReturnDepth": "", "rally_count": 0, "ServeWidth": "", "WinnerType": "0", "WinnerShotType": "0", "shots": {"top": "N. Basilashvili", "bottom": "L. Rosol", "seq": [{"who": "N. Basilashvili", "pos": "top", "stf": 12, "eff": 26, "act": "Serve", "where": "up the middle"}, {"who": "L. Rosol", "pos": "bottom", "stf": 29, "eff": 61, "act": "Backhand", "where": ""}, {"who": "N. Basilashvili", "pos": "top", "stf": 97, "eff": 265, "act": "Serve", "where": "up the middle"}, {"who": "L. Rosol", "pos": "bottom", "stf": 265, "eff": 275, "act": "Forehand", "where": ""}, {"who": "N. Basilashvili", "pos": "top", "stf": 290, "eff": 296, "act": "NotAShot", "where": ""}], "pbp_file_path": "/dccstor/leonidka1/sonia_space/resultsOfMassiveProcessing/videoCVgeneratedMetadata/1105_210_db_metadata.json", {"id": "1115_171_db", "p1": "T. Griekspoor", "p2": "F. Fognini", "p1b": "", "p2b": "", "server": "F. Fognini", "set_num": 3, "point_number": 171, "points": "15-40", "game": "5-2", "description": "Fognini wins the point with an ace.", "excitement_overall_score": 0.3196034629821145, "excitement_crowd_score": 0.5567825938088074, "excitement_analytics_score": 0.1, "excitement_gesture_score": 0.015079165274028783, "winner_of_point": 2, "high_excitement": 0, "long_rally": 0, "winner": 2, "break_point": 0, "break_point_won": 0, "forced_error": 0, "ace": 2, "replay": 0, "volley": 0, "smash": 0, "net_approach": 0, "serve_number": 2, "file_name": "1115_171_db.mp4", "event": "wimbledon", "year": "2022", "server_id": 1, "event_type": "tennis", "day_msg": "Day 1: Mon, Jun 28", "court": "Court 16", "p1_fn": "Tallon Griekspoor", "p2_fn": "Fabio Fognini", "p1b_fn": "", "p2b_fn": "", "p1_nat": "NED", "p2_nat": "ITA", "p1b_nat": "", "p2b_nat": "", "Speed_MPH": 0, "DoubleFault": 0, "UnforcedError": 0, "NetPoint": 0, "BreakPoint": 0, "BreakPointWon": 0, "ServeDepth": "", "ReturnDepth": "", "rally_count": 0, "ServeWidth": "", "WinnerType": "0", "WinnerShotType": "0", "shots": {"top": "F. Fognini", "bottom": "T. Griekspoor", "seq": [{"who": "F. Fognini", "pos": "bottom", "stf": 19, "eff": 45, "act": "Serve", "where": "up the middle"}, {"who": "T. Griekspoor", "pos": "top", "stf": 55, "eff": 92, "act": "Backhand", "where": "up the middle"}, {"who": "F. Fognini", "pos": "bottom", "stf": 127, "eff": 239, "act": "Serve", "where": "up the middle"}, {"who": "T. Griekspoor", "pos": "top", "stf": 253, "eff": 305, "act": "NotAShot", "where": ""}], "pbp_file_path": "/dccstor/leonidka1/sonia_space/resultsOfMassiveProcessing/videoCVgeneratedMetadata/1115_171_db_metadata.json", {"id": "1144_109_db", "p1": "A. Balazs", "p2": "R. Bautista Agut", "p1b": "", "p2b": "", "server": "R. Bautista Agut", "set_num": 3, "point_number": 109, "points": "15-15", "game": "2-1", "description": "Balazs loses the point with a backhand forced error.", "excitement_overall_score": 0.18151347945402357, "excitement_crowd_score": 0.23575143355335748, "excitement_analytics_score": 0.1, "excitement_gesture_score": 0.015001179158329013, "winner_of_point": 2, "high_excitement": 0, "long_rally": 0, "very_long_rally": 0, "winner": 0, "break_point": 0, "break_point_won": 0, "forced_error": 1, "ace": 0, "replay": 0, "volley": 0, "smash": 0, "net_approach": 0, "serve_number": 1, "file_name": "1144_109_db.mp4", "event": "wimbledon", "year": "2022", "server_id": 2, "event_type": "tennis", "day_msg": "Day 2: Tue, Jun 29", "court": "Court 9", "p1_fn": "Attila Balazs", "p2_fn": "Roberto Bautista Agut", "p1b_fn": "", "p2b_fn": "", "p1_nat": "HUN", "p2_nat": "ESP", "p1b_nat": "", "p2b_nat": "", "Speed_MPH": 0, "DoubleFault": 0, "UnforcedError": 0, "NetPoint": 0, "BreakPoint": 0, "BreakPointWon": 0, "ServeDepth": "", "ReturnDepth": "", "rally_count": 0, "ServeWidth": "", "WinnerType": "0", "WinnerShotType": "0", "shots": {"top": "A. Balazs", "bottom": "R. Bautista Agut", "seq": [{"who": "R. Bautista Agut", "pos": "bottom", "stf": 10, "eff": 21, "act": "NotAShot", "where": ""}, {"who": "A. Balazs", "pos": "top", "stf": 20, "eff": 28, "act": "NotAShot", "where": ""}, {"who": "R. Bautista Agut", "pos": "bottom", "stf": 35, "eff": 39, "act": "NotAShot", "where": ""}, {"who": "A. Balazs", "pos": "top", "stf": 40, "eff": 47, "act": "NotAShot", "where": ""}, {"who": "R. Bautista Agut", "pos": "bottom", "stf": 46, "eff": 87, "act": "NotAShot", "where": ""}, {"who": "R. Bautista Agut", "pos": "bottom", "stf": 108, "eff": 124, "act": "Serve", "where": "up the middle"}, {"who": "A. Balazs", "pos": "top", "stf": 133, "eff": 140, "act": "NotAShot", "where": ""}], "pbp_file_path": "/dccstor/leonidka1/sonia_space/resultsOfMassiveProcessing/videoCVgeneratedMetadata/1144_109_db_metadata.json", {"id": "1219_21_db", "p1": "A. De Minaur", "p2": "J. Draper", "p1b": "", "p2b": "", "server": "J. Draper", "set_num": 1, "point_number": 21, "points": "15-40", "game": "1-2", "description": "Draper wins the point with an ace.", "excitement_overall_score": 0.3127906036660247, "excitement_crowd_score": 0.2544923735988407, "excitement_analytics_score": 0.1, "excitement_gesture_score": 0.015133160588898463, "winner_of_point": 2, "high_excitement": 0, "long_rally": 0, "very_long_rally": 0, "winner": 2, "break_point": 0, "break_point_won": 0, "forced_error": 0, "ace": 2, "replay": 0, "volley": 0, "smash": 0, "net_approach": 0, "serve_number": 1, "file_name": "1219_21_db.mp4", "event": "wimbledon", "year": "2022", "server_id": 2, "event_type": "tennis", "day_msg": "Day 4: Thur, Jul 1", "court": "No.1 Court", "p1_fn": "Alex De Minaur", "p2_fn": "Jack Draper", "p1b_fn": "", "p2b_fn": "", "p1_nat": "AUS", "p2_nat": "GBR", "p1b_nat": "", "p2b_nat": "", "Speed_MPH": 131, "DoubleFault": 0, "UnforcedError": 0, "NetPoint": 0, "BreakPoint": 0, "BreakPointWon": 0, "ServeDepth": "CTL", "ReturnDepth": "", "rally_count": 1, "ServeWidth": "C", "WinnerType": "0", "WinnerShotType": "0", "shots": {"top": "A. De Minaur", "bottom": "J. Draper", "seq": [{"who": "J. Draper", "pos": "bottom", "stf": 0, "eff": 25, "act": "Serve", "where": "up the middle"}, {"who": "A. De Minaur", "pos": "top", "stf": 37, "eff": 63, "act": "Serve", "where": ""}], "pbp_file_path": "/dccstor/leonidka1/sonia_space/resultsOfMassiveProcessing/videoCVgeneratedMetadata/1219_21_db_metadata.json", {"id": "1107_106_db", "p1": "T. Van Rijthoven", "p2": "F. Delbonis", "p1b": "", "p2b": "", "server": "F. Delbonis", "set_num": 2, "point_number": 106,
```

Our Layered LoRA Architecture (LaLo)



LaLo can support task modularity & efficient recall of past models

Efficient adaptation: < 3% of all parameters are adapted !

Wimbledon Multimodal Perception and Commentary

IBM

All Events

All Rounds

All Statistics

Day 3 Wed 3 July

Type Player Name

COVERAGE

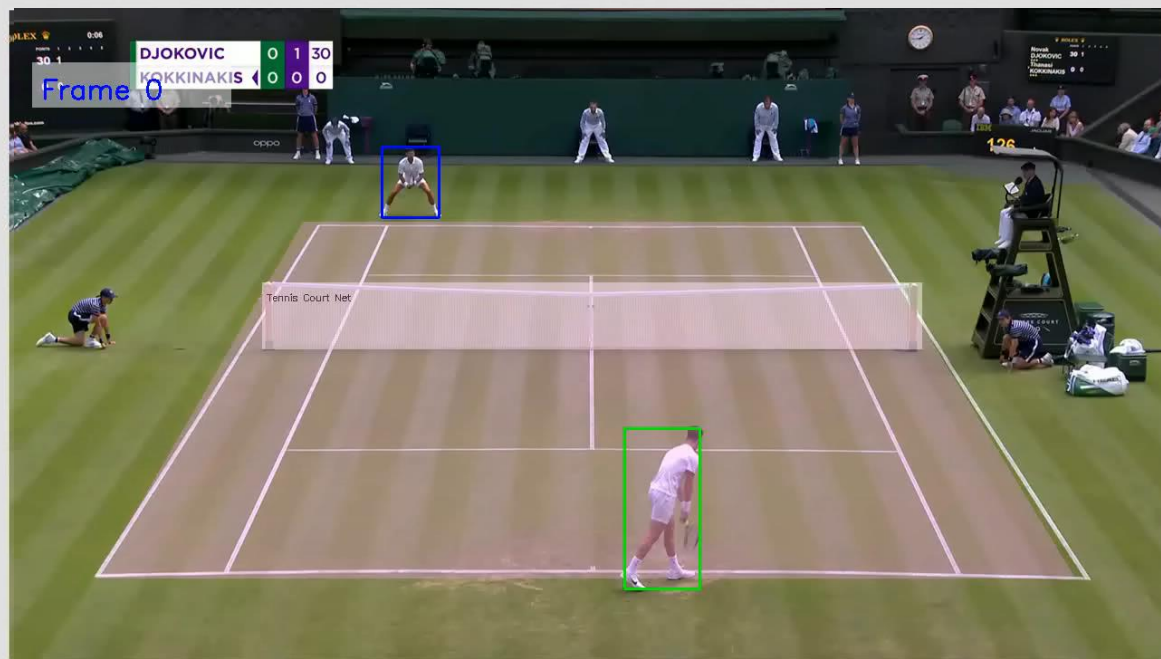
13169 48
Points Hours

HIGHLIGHTS

14 0
Produced Published

TOP 5 MAIN EVENTS HIGHLIGHTS

0 0 900
Published Secs Published Secs Remaining



Djokovic vs Kokkinakis

Set 1: 1-0; 15-30

Auto-Generated Commentary:

After a Kokkinakis serve up the middle, Djokovic hits a backhand. After a Kokkinakis forehand up the middle, the athletes jump into a sequence of backhand to backhand. Unlucky miss by Djokovic as he directs a backhand unforced error into the net to lose the point.

Most Excitement | Most Recent

Set Excitement Threshold

0.00 0.3 1.00



0.90

Wednesday, 3 July 2019, 13:08

Novak Djokovic vs. Thanasi Kokkinakis

Set 1: 15-30 : Djokovic loses the point with a backhand unforced error.



0.15

Wednesday, 3 July 2019, 13:08

Serena Williams vs. Alize Cornet

Set 3: 15-40 : Break Point; Williams wins the point with a forehand winner.



0.27

Wednesday, 3 July 2019, 13:07

Novak Djokovic vs. Thanasi Kokkinakis

Set 1: 0-30 : Kokkinakis loses the point with a forehand unforced error.



0.86

Wednesday, 3 July 2019, 13:06

Jeremy Chardy vs David Goffin

Set 1: 15-40 : Break Point; Goffin wins the point with a forehand winner.

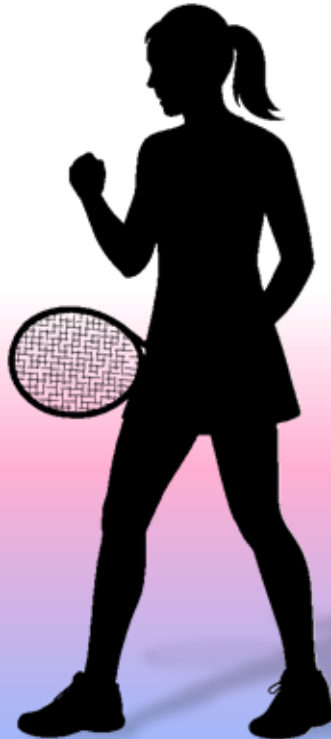
How about voice-over narration?

Excitement-Driven AI Commentary

Low Excitement



Moderate Excitement



High Excitement



Excitement-Driven AI Commentary

Step 1: Use Granite Vision (and audio) to describe and measure the excitement level of the crowd and player reaction

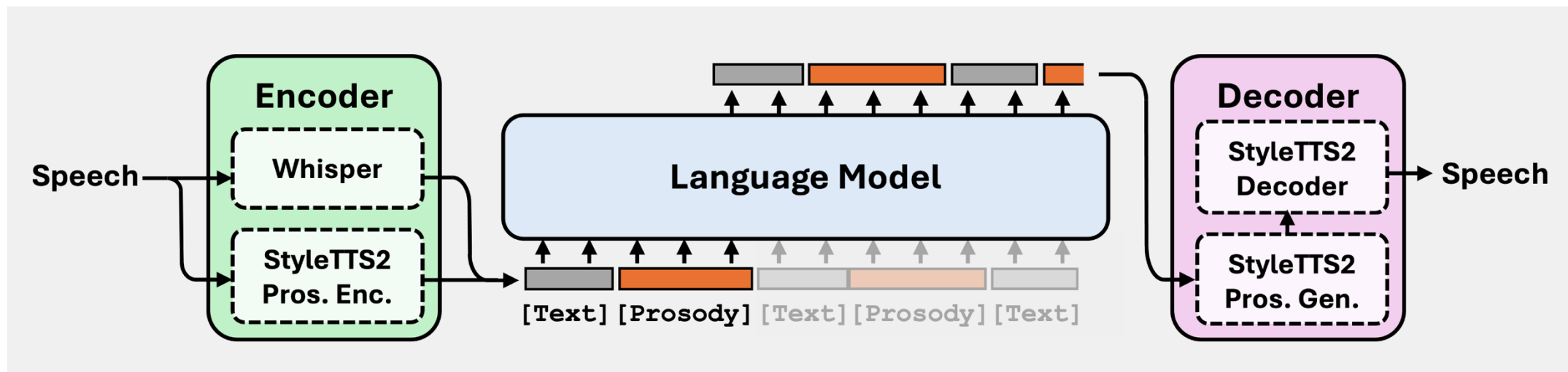


Step 2: Adapt the content and prosody/tone of AI commentary based on the excitement level

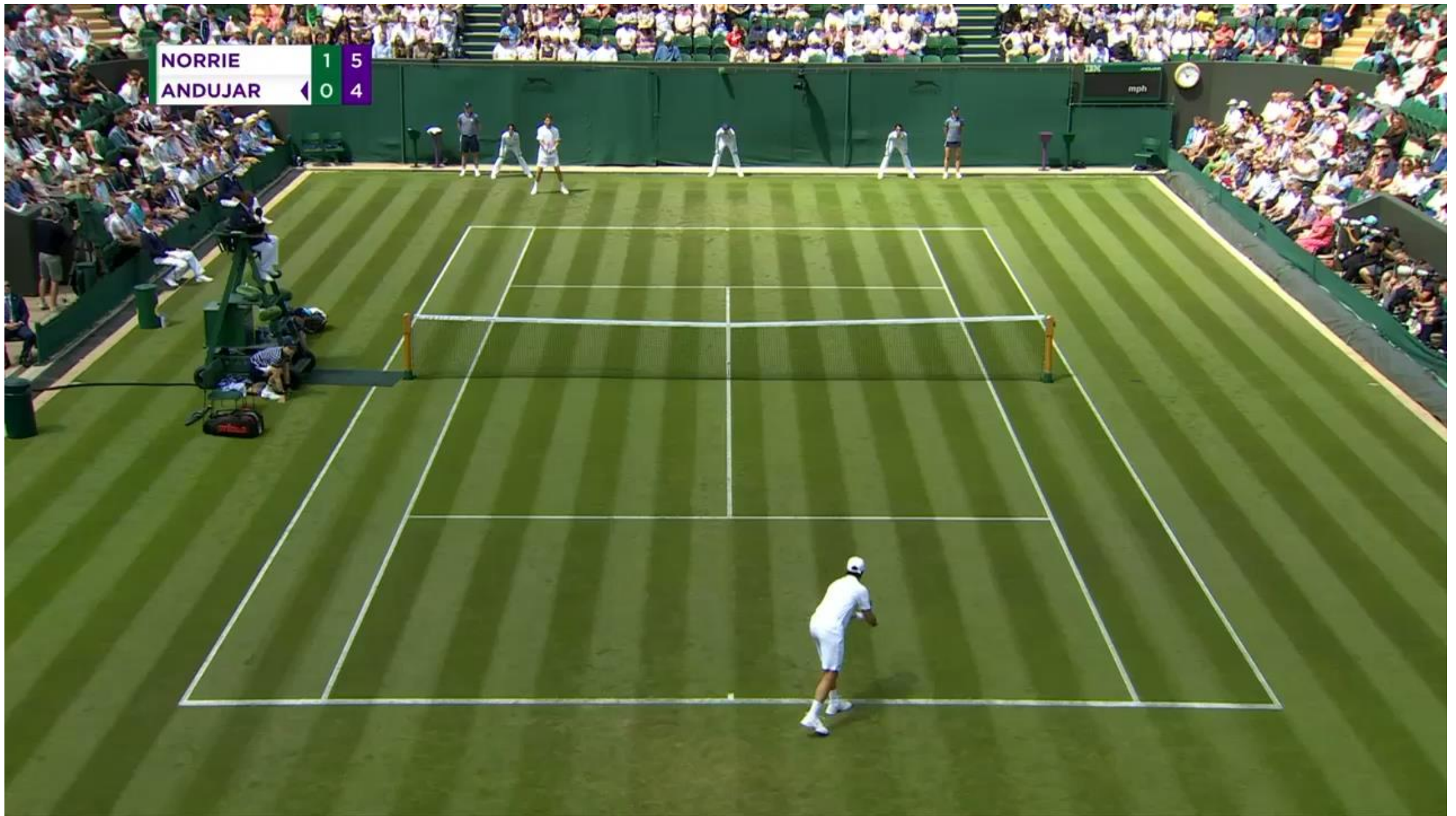
→ ProsodyLM [Kaizhi Qian, Yang Zhang et al, COLM 2025]

Controlling Expressive Speech Generation

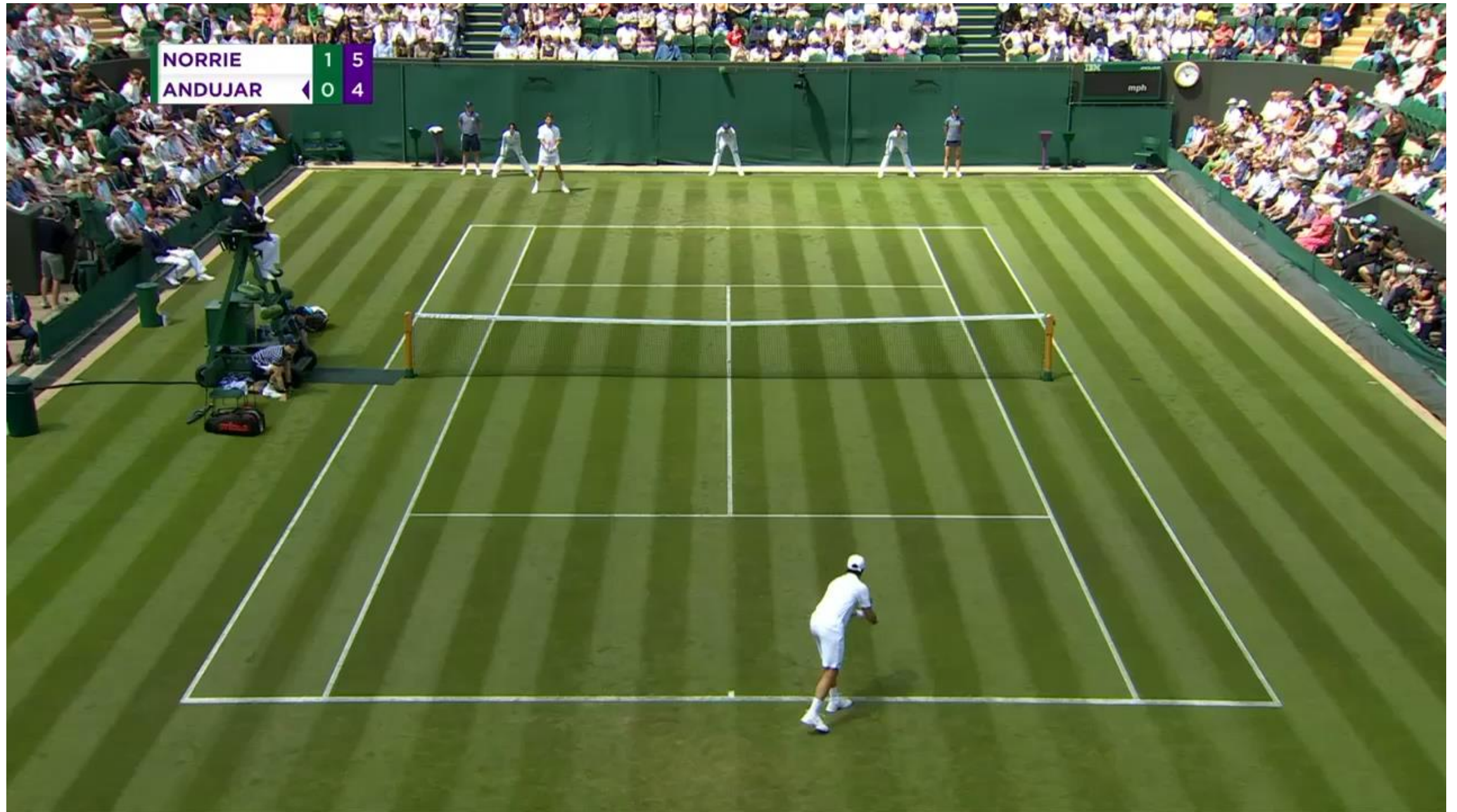
ProsodyLM [Kaizhi Qian, Yang Zhang et al, COLM 2025]



Low Excitement



High Excitement



MATCH POINT

26	BADOSA	ESP		2	5	30
13	NAVARRO	USA		6	6	40



Ongoing Research: Streaming VLM + Multimodal ProsodyLM

- Predicting Text+Prosody from Video Tokens
- Real-time, streaming live commentary
- Personalized prosody control

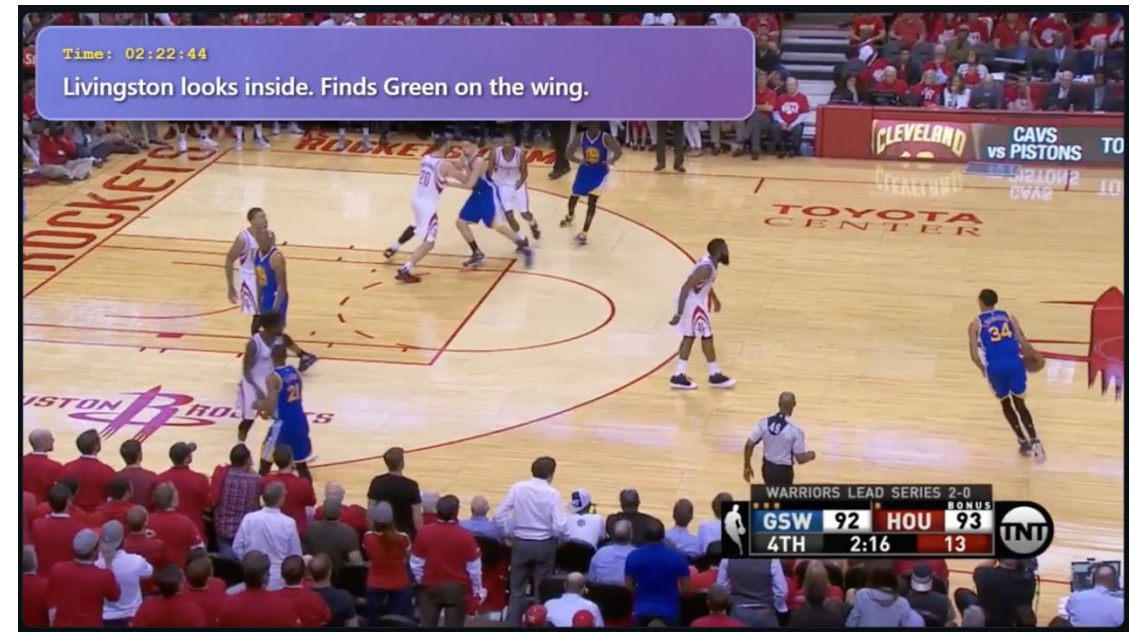
[cs.CV] 10 Oct 2025

STREAMINGVLM: REAL-TIME UNDERSTANDING FOR INFINITE VIDEO STREAMS

Ruyi Xu^{1*} Guangxuan Xiao^{1*} Yukang Chen² Liuning He¹
Kelly Peng³ Yao Lu² Song Han^{1,2}

¹MIT ²NVIDIA ³First Intelligence

<https://github.com/mit-han-lab/streaming-vlm>



Thank you!

(see more at <http://rogerioferis.org>)