VIVA: An End-to-End System for Interactive Video Analytics

Francisco Romero*, Daniel Kang*, Johann Hauswald, Ioanna Vavelidou, Peter Bailis, Christos Kozyrakis, Matei Zaharia
Growing demand for video analytics

“How many cars passed by on Monday?”

“Find clips of Jake Tapper interviewing an angry Bernie Sanders”

“Show me Ronaldo’s headers when he played for Juventus”
Example – analyzing TV news

- Detect angry faces in frames
- Detect Sanders from angry face frames
- Detect Tapper in frames with “angry Sanders”

- Train specialized model to detect Tapper and Sanders
- Detect frames with Tapper and Sanders
- Detect “angry Sanders” frames

Query result

Not obvious which is best!

“Find clips of Jake Tapper interviewing an angry Bernie Sanders”
Existing work focuses on specific components

<table>
<thead>
<tr>
<th>System</th>
<th>Component</th>
<th>Query</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoScope, Focus, ...</td>
<td>Query optimizer/execution</td>
<td>Selection</td>
</tr>
<tr>
<td>MIRIS</td>
<td>Query execution</td>
<td>Tracking</td>
</tr>
<tr>
<td>Blazelt</td>
<td>Query optimizer</td>
<td>Aggregation, limit</td>
</tr>
<tr>
<td>TASTI</td>
<td>Index</td>
<td>Proxy-based</td>
</tr>
</tbody>
</table>

Users today must manually combine these techniques/systems!
Video query challenges

**Goal**: enable videos to be searched like structured data

**Challenges**: 
- Ad-hoc, exploratory queries
- Require expensive DNNs to extract information
  - Expensive hardware: A100 is $4.10/hr
  - Slow: as low as 3 frames per second
- Optimization space spans structured and unstructured data
VIVA

» End-to-end interactive video analytics system

» Enables interactive exploration of large video corpuses by:
  » Allowing users to describe domain-specific relations
  » Blending structured and unstructured data during query optimization
  » Co-optimizing storage and compute for efficient DNN execution

Under active development!

SELECT time_window FROM news_analysis
WHERE tapper_angry_sanders = TRUE
VIVA workflow – analyzing TV news

<table>
<thead>
<tr>
<th>news_analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>video_id</td>
</tr>
</tbody>
</table>

Table: news_analysis

<table>
<thead>
<tr>
<th>video_id</th>
<th>time_window</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
VIVA workflow – analyzing TV news

```
ALTER TABLE news_analysis ADD face_label
AS face_detection(frame)

-- Add other query-specific columns
```

<table>
<thead>
<tr>
<th>video_id</th>
<th>time_window</th>
<th>face_label</th>
<th>sentence_segment</th>
<th>sanders_frame</th>
<th>...</th>
<th>tapper_angry_sanders</th>
</tr>
</thead>
</table>

How does VIVA execute video queries interactively and at low-cost?

ALTER TABLE news_analysis ADD face_label
AS face_detection(frame)

--Add other query-specific columns

SELECT time_window FROM news_analysis
WHERE tapper_angry_sanders = TRUE
Optimizing query plan execution

» **Mixed-data optimizer:**
  » Optimizes across structured and unstructured data
  » Accounts for heterogeneous hardware backends for unstructured data

» **Accelerator-based execution engine:**
  » Makes scheduling/resource allocation decisions for decoding, DNN computation, and transcoding
Optimizing video fetch and caching results

» **Video file manager:**
  » Per-compute instance
  » Pre-fetches, caches, and manages video data

» **Embedding cache:**
  » Decides which embeddings to store
  » Tracks DNN version that produced embedding
VIVA workflow – analyzing TV news

ALTER TABLE news_analysis ADD face_label AS face_detection(frame)

--Add other query-specific columns

SELECT time_window FROM news_analysis
WHERE tapper_angry_sanders = TRUE

<table>
<thead>
<tr>
<th>video_id</th>
<th>time_window</th>
<th>face_label</th>
<th>sentence_segment</th>
<th>sanders_frame</th>
<th>...</th>
<th>tapper_angry_sanders</th>
</tr>
</thead>
<tbody>
<tr>
<td>video_1</td>
<td>[1,5]</td>
<td>Sanders</td>
<td>“Tapper: Senator Sanders, why are you upset...”</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>video_1</td>
<td>[1,5]</td>
<td>Tapper</td>
<td>“Tapper: Senator Sanders, why are you upset...”</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>video_1</td>
<td>[55,60]</td>
<td>Sanders</td>
<td>“Bush: Senator Sanders has an interesting point...”</td>
<td>0</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

![Video analysis diagram](image-url)
VIVA workflow – analyzing TV news

ALTER TABLE news_analysis ADD face_label AS face_detection(frame)

--Add other query-specific columns

SELECT time_window FROM news_analysis
WHERE tapper_angry_sanders = TRUE

ALTER TABLE news_analysis ADD sanders_transcript AS sentence_segment LIKE '%Sanders%'

SELECT time_window FROM news_analysis
WHERE sanders_transcript = TRUE

<table>
<thead>
<tr>
<th>video_id</th>
<th>time_window</th>
<th>face_label</th>
<th>sentence_segment</th>
<th>sanders_frame</th>
<th>sanders_transcript</th>
<th>...</th>
<th>tapper_angry_sanders</th>
</tr>
</thead>
<tbody>
<tr>
<td>video_1</td>
<td>[1,5]</td>
<td>Sanders</td>
<td>“Tapper: Senator Sanders, why are you upset...”</td>
<td>1</td>
<td>1</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>video_1</td>
<td>[1,5]</td>
<td>Tapper</td>
<td>“Tapper: Senator Sanders, why are you upset...”</td>
<td>1</td>
<td>1</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>video_1</td>
<td>[55,60]</td>
<td>Sanders</td>
<td>“Bush: Senator Sanders has an interesting point...”</td>
<td>0</td>
<td>1</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
VIVA workflow – analyzing TV news

```
ALTER TABLE news_analysis ADD face_label
AS face_detection(frame)

ALTER TABLE news_analysis ADD sanders_transcript
AS sentence_segment LIKE ‘%Sanders%’

SELECT time_window FROM news_analysis
WHERE tapper_angry_sanders = TRUE

SELECT time_window FROM news_analysis
WHERE sanders_transcript = TRUE
```

<table>
<thead>
<tr>
<th>video_id</th>
<th>time_window</th>
<th>face_label</th>
<th>sentence_segment</th>
<th>sanders_frame</th>
<th>sanders_transcript</th>
<th>...</th>
<th>tapper_angry_sanders</th>
</tr>
</thead>
<tbody>
<tr>
<td>video_1</td>
<td>[1,3]</td>
<td>Sanders</td>
<td>“Tapper: Senator Sanders, why are you upset...”</td>
<td>1</td>
<td>1</td>
<td>...</td>
<td>1</td>
</tr>
</tbody>
</table>

| video_1  | [1,3]       | Tapper     | “Tapper: Senator Sanders, why are you upset...” | 1             | 1                   | ... | 1                   |

| video_1  | [55,60]     | Sanders    | “Bush: Senator Sanders has an interesting point...” | 0             | 1                   | ... | 0                   |

How can users express query- and domain-knowledge about virtual columns to VIVA?
Specifying domain knowledge

- **Relational hints explorer**: explore and declare virtual column relationships
  - `columnA SUPERSET columnB`
    - Sanders in frame SUPERSET Sanders/Tapper interviews
  - `columnA EQUALS columnB`
    - Sanders in frame EQUALS Sanders in transcript
  - `columnA EXCLUDES columnB`
    - Sanders/Tapper interviews EXCLUDES Cooking shows
  - `columnA PROXY columnB`
    - People wearing glasses PROXY Sanders in frame
VIVA workflow – analyzing TV news

ALTER TABLE news_analysis ADD face_label AS face_detection(frame)

--Add other query-specific columns

SELECT time_window FROM news_analysis WHERE tapper_angry_sanders = TRUE

ALTER TABLE news_analysis ADD sanders_transcript AS sentence_segment LIKE '%Sanders%'

SELECT time_window FROM news_analysis WHERE sanders_transcript = TRUE

CREATE HINT transcript_frame_equals FOR news_analysis AS sanders_frame EQUALS sanders_transcript

<table>
<thead>
<tr>
<th>video_id</th>
<th>time_window</th>
<th>face_label</th>
<th>sentence_segment</th>
<th>sanders_frame</th>
<th>sanders_transcript</th>
<th>...</th>
<th>tapper_angry_sanders</th>
</tr>
</thead>
<tbody>
<tr>
<td>video_1</td>
<td>[1,3]</td>
<td>Sanders</td>
<td>“Tapper: Senator Sanders, why are you upset...”</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>video_1</td>
<td>[1,3]</td>
<td>Tapper</td>
<td>“Tapper: Senator Sanders, why are you upset...”</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>video_1</td>
<td>[55,60]</td>
<td>Sanders</td>
<td>“Bush: Senator Sanders has an interesting point...”</td>
<td>0</td>
<td>1</td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

...
-- Add other query-specific columns

**ALTER TABLE** news_analysis **ADD** face_label
**AS** face_detection(frame)

**SELECT** time_window **FROM** news_analysis
**WHERE** tapper_angry_sanders = TRUE

**ALTER TABLE** news_analysis **ADD** sanders_transcript
**AS** sentence_segment **LIKE** '%Sanders%'

**SELECT** time_window **FROM** news_analysis
**WHERE** sanders_transcript = TRUE

**CREATE HINT** transcript_frame_equals
**FOR** news_analysis
**AS** sanders_frame **EQUALS** sanders_transcript

**SELECT** time_window **EQUALS** sanders_transcript
**WHERE** tapper_angry_sanders = TRUE
**USING HINT** transcript_frame_equals

<table>
<thead>
<tr>
<th>video_id</th>
<th>time_window</th>
<th>face_label</th>
<th>sentence_segment</th>
<th>sanders_frame</th>
<th>sanders_transcript</th>
<th>...</th>
<th>tapper_angry_sanders</th>
</tr>
</thead>
<tbody>
<tr>
<td>video_1</td>
<td>[1,3]</td>
<td>Sanders</td>
<td>“Tapper: Senator Sanders, why are you upset...”</td>
<td>1</td>
<td>1</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>video_1</td>
<td>[1,3]</td>
<td>Tapper</td>
<td>“Tapper: Senator Sanders, why are you upset...”</td>
<td>1</td>
<td>1</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>video_1</td>
<td>[55,60]</td>
<td>Sanders</td>
<td>“Bush: Senator Sanders has an interesting point...”</td>
<td>1</td>
<td>1</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>video_8320</td>
<td>[566, 569]</td>
<td>Tapper</td>
<td>“Tapper: That’s all the time we have for today.”</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
</tbody>
</table>

---

VIVA workflow – analyzing TV news
Open research directions

» Automatically inferring relational hints
» Determining when DNNs should be (re)-trained
» Infrastructure for high throughput demands
Conclusion

» End-to-end interactive video analytics system
» Enable users to search videos as easily as structured data
» Under active development

Questions?
faromero@stanford.edu