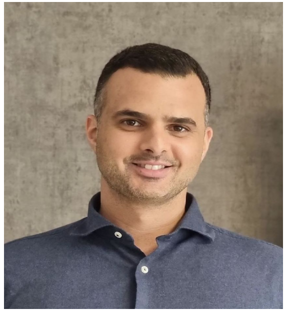
Lightricks

hyadata lab

THE HEBREW UNIVERSITY OF JERUSALEM

🎥 🔧 🔗 Visual Editing with LLM-based Tool Chaining: An Efficient Distillation Approach for Real-Time Applications


Oren Sultan


Alex Khasin


Guy Shiran


Asi Messica


Prof. Dafna Shahaf

EMNLP 2024

Background and Motivation

- Videos are a popular storytelling medium; however, the intricate nature of video editing poses substantial challenges for novice users.

- Using natural language can mitigate this challenge – text-to-video, diffusion-based models have demonstrated impressive results. However, they are computationally expensive, slow, and still lack in visual quality and user control over the generated video.

- We believe it is better to teach LLMs to use specialized tools than rely on black-box models.

Background and Motivation

- **<u>Idea.</u>** to teach LLMs to use existing, **specialized** tools in **VideoLeap**

- **<u>Goal.</u>** to implement an AI assistant, democratizing advanced capabilities.

- As a **proof-of-concept**, we focused on **tonal color adjustments**, allowing users to change a video's appearance via textual instructions.

Visual Editing Example

Adjust

{
"exposure": 0,
"contrast": 10,
"brightness": 10,
"highlights": 20,
"shadows": -10,
"saturation": 15,
"vibrance": 15,
"temperature": 30,
"tint": 10,
"hue": 0,
"bloom": 0,
"sharpen": 0,
"structure": 0,
"linearOffset": 0
}

Selective adjust

{
"red": {"saturation": 20, "luminance": 10},
"orange": {"saturation": 30, "luminance": 20},
"yellow": {"saturation": 40, "luminance": 30},
"green": {"saturation": -20, "luminance": 0},
"cyan": {"saturation": -20, "luminance": 0},
"blue": {"saturation": 0, "luminance": 0}
}

Filter

{
"name": "faded_HighNoon",
"intensity": 40
}

"Golden hour"

Proof-of-concept with GPT-3.5-Turbo

Current drawbacks

- Dependency on GPT-3.5-Turbo, a **closed model** with usage **costs**
- Larger LMs like GPT-3.5-Turbo have **high latency**
- Lack of integration of user preferences

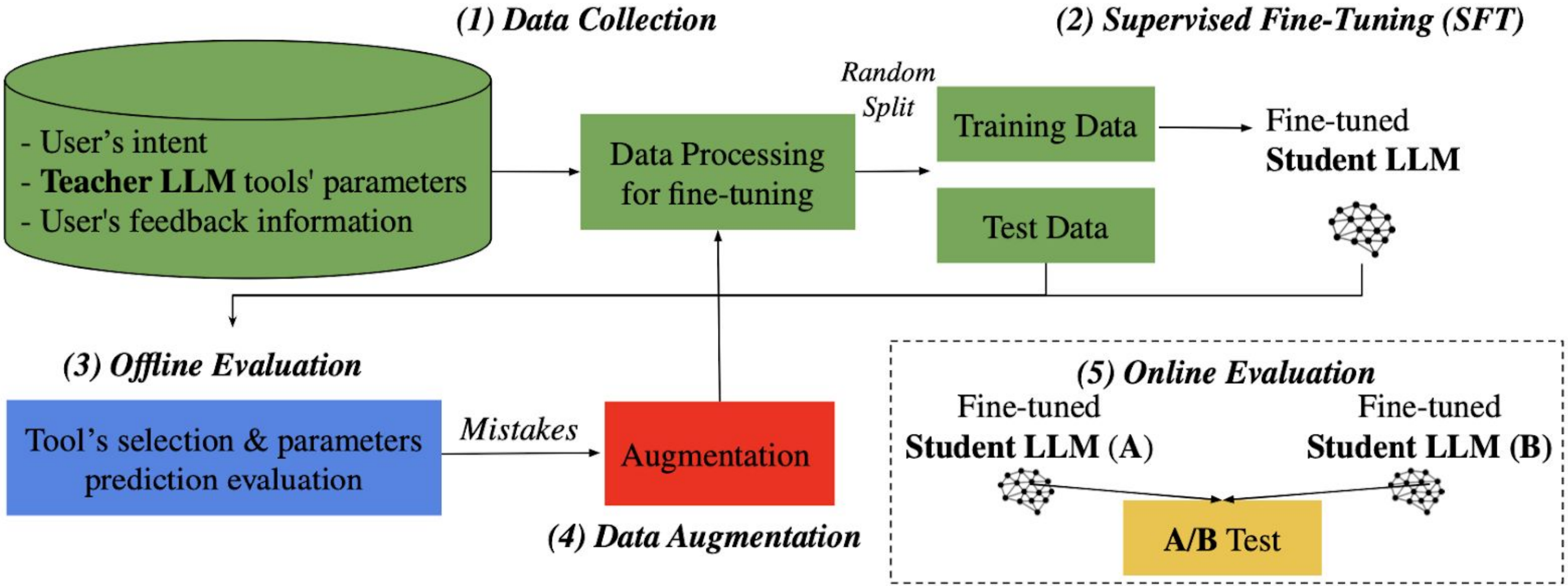
Our proposed solution

A Distillation framework – fine-tune a (**smaller**) **student LLM** with guidance from a (**larger**) **teacher LLM** and **users behavioral signals**

Our proposed solution advantages

- Open-source models are **free**
- Smaller LMs have a **better latency**
- Fine-tuning on high-quality data to better **align our user preferences**

Our distillation framework approach



Offline Evaluation Metrics

- **Tool-selection:** the model's ability to decide correctly whether to use a tool.
  We measure ***precision*** and ***recall***, and report tool-selection score as the ***F1-score***.

- **Quality:** the model's ability to use a tool correctly.
  - For the **filter tool**: the ***accuracy*** on the filter name.
  - For the **adjust** and **selective adjust** tools: the ***mean cosine similarity*** across samples between predicted and ground-truth parameter values.

- **Final score:** the ***harmonic mean*** between ***tool-selection score*** and ***quality score***, emphasizing high performance in both.

- **Overall score:** the average of the final scores of all tools.

- **Reality check** on the generated images/videos.

Online Evaluation

- When our offline evaluation shows it is worthwhile to consider a new student LLM, we confirm it in an online A/B test experiment.

- **Metric of interest**: *project_completion_rate = #projects_exported / #projects_started*.

- This metric indicates the total user satisfaction with the results and the overall experience.

Experiments

Research Questions.

- **RQ1:** How well do student LLMs perform, and do they effectively mimic the teacher LLM?
- **RQ2:** Is augmentation effective in low-data regimes?

Models.

- **Teacher LLM**: **GPT-3.5-Turbo**
- **Student LLMs**:
  - **Llama-2-7b-chat-hf** with Low Rank Adaptations (LoRA) + Quantization, A100 GPU.
  - **FlanT5-base (250M)** (faster), L4 GPU (5 times cheaper).

RQ1: Student LLMs Performance – Offline Evaluation

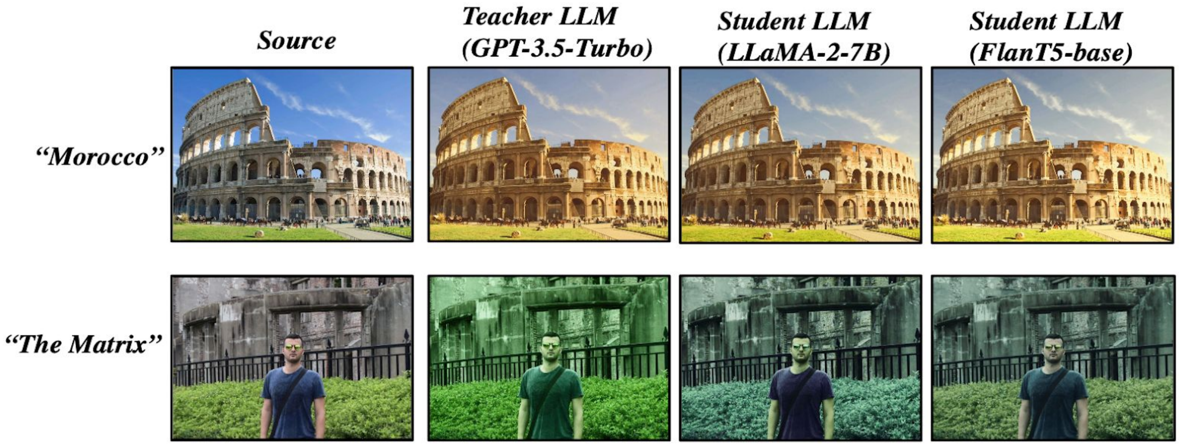
Row	Model	Test	Adjust	Selective Adjust	Filter	Overall
1	Llama-2-7b-chat-hf	All	(.95, .63, .76)	(.75, .66, .70)	(.81, .71, .76)	.74
2		r_3	(.98, .68, .80)	(.82, .67, .74)	(.92, .73, .81)	.78
3		r_5	(.98, .75, .85)	(.87, .71, .78)	(.91, .83, .87)	.83
4	FlanT5-base (250M)	All	(.95, .57, .72)	(.76, .65, .70)	(.78, .71, .74)	.72
5		r_3	(.99, .61, .76)	(.87, .66, .75)	(.88, .72, .79)	.77
6		r_5	(.99, .68, .80)	(.90, .71, .79)	(.89, .82, .85)	.81

- **Metrics**: (tool-selection score, quality score, final score).
  - **Overall**: average of final scores across the tools.

- **FlanT5-base performs very similarly to Llama-2-7b-chat-hf (rows 1, 4)**.

RQ1: Student LLMs Performance – Offline Evaluation

- **Reality check** – human manual annotation on a sample of 15 generated images.
- Three calibrated team annotators reviewed each sample according to two criteria:
  - Is the image relevant to the intent?
  - Does the student model correctly mimic the teacher?



- **Relevancy**: 13-14 out of 15 for all models.
- **Student LLM correctly mimic the teacher**: 11 out of 15 for both (not the same).

RQ1: Student LLMs Performance – Online Evaluation (A/B Test)

Metric. project completion rate (as an indicator for user satisfaction)

Experiment 1.

- **Teacher LLM:** GPT-3.5-Turbo **vs. Student LLM:** Llama-2-7b-chat
- **Conclusion:** Similar performance, we chose Llama-2-7b-chat for its lower latency and cost

Experiment 2.

- **Student LLM:** FlanT5-base **vs. Student LLM:** Llama-2-7b-chat
- **Conclusion:** Similar performance, we chose FlanT5-base for its lower latency and cost

- Our offline metrics align with the results of the online A/B tests.

RQ2: Augmentation in low-data regimes

- We show a 25% improvement in fine-tuning in low-data regimes using data augmentation

Train %	Augmentations	Train Size	Overall Score
100	0	8,252	0.72
12.5%	0	1,031	0.52
12.5%	**806 (43.8%)**	**1,837**	**0.65**

Conclusions

- We presented a novel NLP application for **automatic video editing using LLMs**, focusing on **tonal color adjustment**.

- By fine-tuning a **student LLM** with guidance from a **larger teacher LLM** and **user behavioral signals,** we achieved similar performance to GPT-3.5-Turbo both in **offline and online experiments**.

- Our solution significantly **reduces costs and latency**, crucial for industry applications.

- **Paper website**: https://www.orensultan.com/ai_recolor.github.io/
- See you in Miami! 🇺🇸