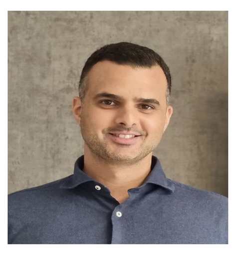
Lightricks

האוניברסיטה העברית בירושלים
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


Dafna Shahaf

Background and Motivation

- Videos are a popular storytelling medium; however, the intricate nature of video editing poses substantial challenges for novice users.
- Natural language video editing can mitigate this challenge, but current text-to-video models are too slow, costly, and lack quality.
- We believe it's better to teach LLMs to use specialized tools than rely on black-box models. This approach is also more interpretable.

- **Idea:** to teach LLMs to use existing, **specialized tools** in **VideoLeap**
- **Goal:** To implement an AI assistant, democratizing advanced capabilities.
- **Proof-of-concept:** **tonal color adjustments**, allowing users to change a video's appearance via textual instructions.

Our Task

Users provide an image/video and describe the desired appearance.
An LLM interprets the request, selects tools, and sets parameters.
The bottom row shows generated images by applying the LLM's output in our app.

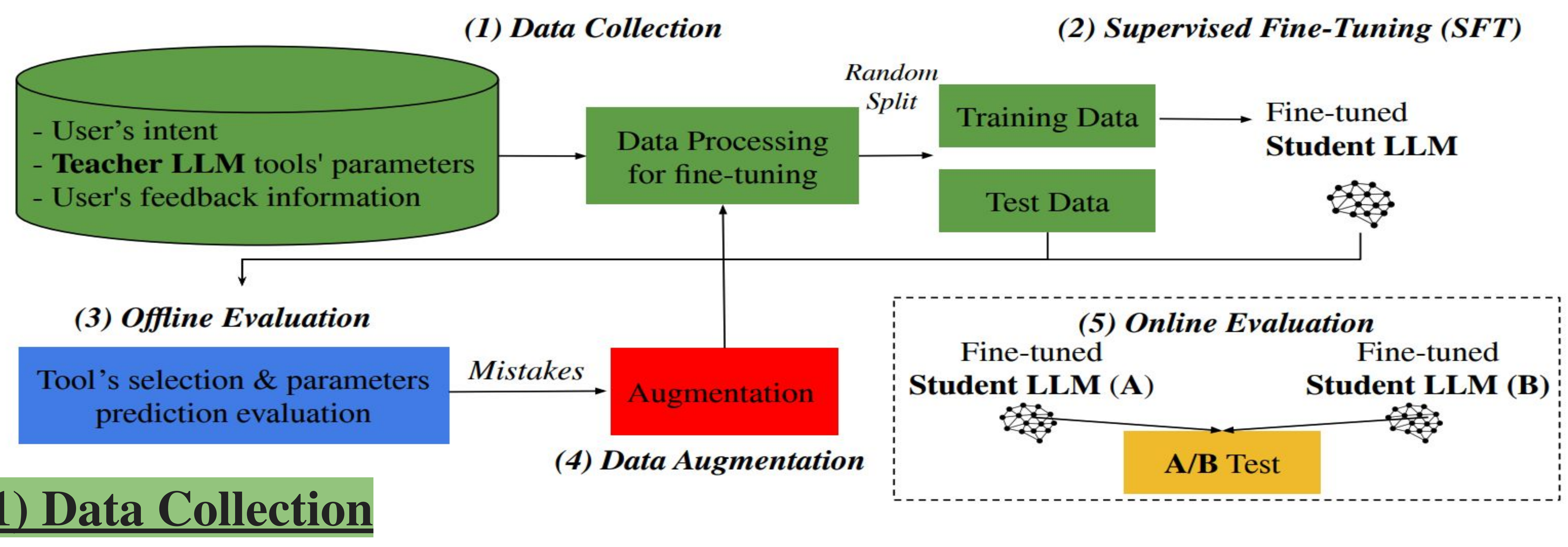
Example: **"Golden hour"**

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}

Our Distillation Framework Approach



(1) Data Collection

Gathering Teacher LLM Outputs

- **Teacher LLM: GPT-3.5-Turbo** (**four months** – data collection period).
- **A data row includes: user's intent**, **output of the teacher LLM** (tools to use, parameters and their values), whether the **user exports** the result per tool.
- **Data Filtering:** samples with zero exports. Our teacher LLM can generate different outputs per intent (across different calls); We take as ground truth the result that **maximizes the export rate**.
- **Prompts: one-shot** example for user intent, with **rational (CoT)** and **output parameters** per tool.
- In total, we collected **9,252 unique user intents**, resulting in **27,756 rows**.

Data Processing for Fine-Tuning

- We used the collected data to fine-tune a student LLM (**more concise** prompts).
- We don't request rational from the student, as we **prioritize low latency**.
- The student LLM is **trained on all three tools** (similar to multi-task instruction).

Data Splitting

- **Test set: 1K unique user intents**, each with a teacher LLM output for each tool.
- **Training set:** the remaining data (**8,252 rows**).
- Each row includes a user intent and three tool outputs.

Set	Adjust		SelectiveAdjust		Filter	
	Used	All	Used	All	Used	All
Train	7570	8252	2647	8252	5448	8252
Test	912	1000	356	1000	683	1000

(2) Supervised Fine-tuning (SFT)

Student LLMs

- **Auto-regressive model (decoder only)**: Llama-2-7b-chat-hf (7B) Meta
- **Sequence-to-Sequence model (encoder decoder)**: FlanT5-base (250M)

(3) Offline Evaluation Metrics

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

- We analyze the actual generated images/videos by applying the tools' predicted parameters in our app.
- We analyze a random sample, with three human annotators per sample (RQ1).
- Ideas for automatic evaluation of the generated images/videos.

(4) Data Augmentation

- We iteratively run the offline evaluation on the training set.
- **(1) Identifying where the student LLM predictions differ from the teacher's**
  - For the **filter tool**, a mistake occurs when the predicted filter name is incorrect.
  - For the **adjust and selective adjust**, a mistake occurs when a sample's cosine similarity is lower than the tool's mean cosine similarity without augmentation.
- **(2) Using another LLM to generate similar input user intents where the student LLM made mistakes (e.g., "cool tone" from "cool morning")**
  - The new intents and the teacher LLM's original answers are added to the training
  - We augmented an intent whenever a mistake was identified by at least one tool.

(5) Online Evaluation (A/B test)

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

Experiments

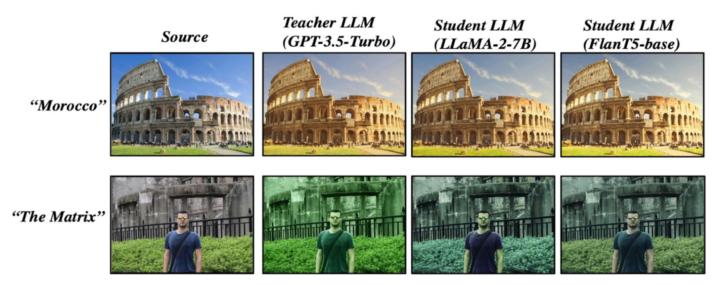
RQ1: How do student LLMs perform, do they effectively mimic the teacher LLM?

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**: avg. of final scores across the tools.
- **FlanT5-base performs very similarly to Llama-2-7b-chat-hf (rows 1, 4)**!

Reality check – human 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).

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

- **Experiment 1. Teacher LLM:** GPT-3.5-Turbo **vs. Student LLM:** Llama-2-7b-chat
  - **Results:** the completion rate for the teacher was **96.1%** of that of Llama-2-7b-chat.
  - 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
  - **Results:** the completion rate of FlanT5-base was **99%** of that of Llama-2-7b-chat.
  - We chose FlanT5-base for its lower latency and cost.

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

RQ2: Is augmentation effective in low-data?

25% performance improvement (+0.13),
in low data regimes (1/8 of the training)
with just one iteration!

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**

Future Work

- To test potential fine-tuning improvements by **adding rational as an additional label** for supplementary supervision in a **multi-task framework** (Hsieh et al., 2023).
- To quantify the **benefits of integrating user signals**, and to explore **other methods for combining user feedback** (e.g, **personalization**).
- To extend our **one-hop responses** to **conversational agents / dialogue systems**.
- To apply our research into additional **tools, features, and applications**.