KDD2024
BARCELONA, SPAIN

Alibaba Tongyi

Multi-modal Data Processing for Foundation Models: Practical Guidances and Use Cases


Daoyuan Chen


Yaliang Li


Bolin Ding

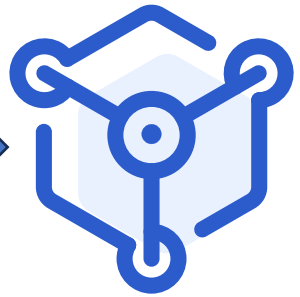
Data-Juicer Team

Tongyi Lab, Alibaba Group

GitHub: https://github.com/modelscope/data-juicer

Multi-modal Foundation Model

Pre-training



Post-tuning

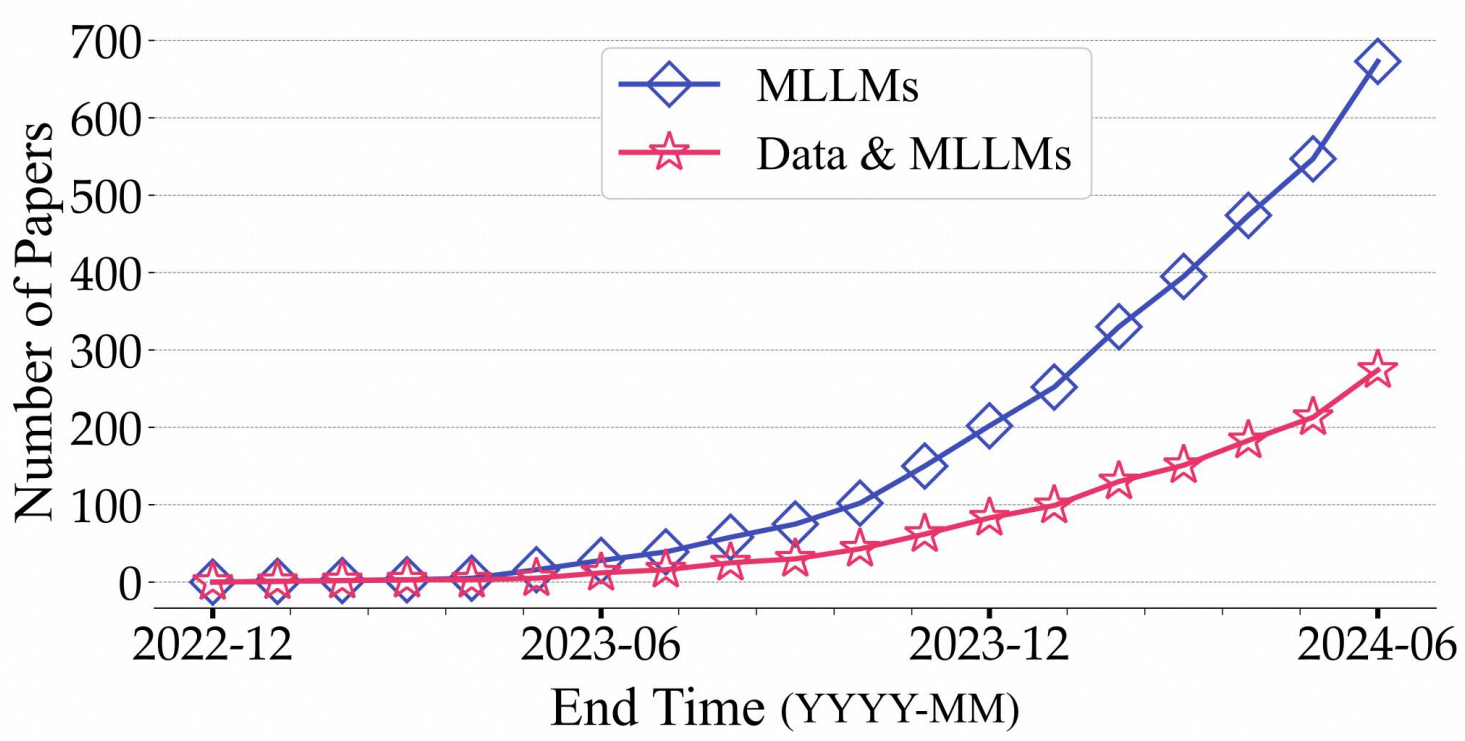
Applications

- AI Generative Content
- Multi-modal Dialogue
- Film Editor
- Education
- Entertainment
- Media
- ……

Multi-modal Data
Video, Image, Text, Voice, ……

Multi-modal Large Language Models (MLLMs)
GPT4v, Sora, Qwen-VL, Gemini, ……

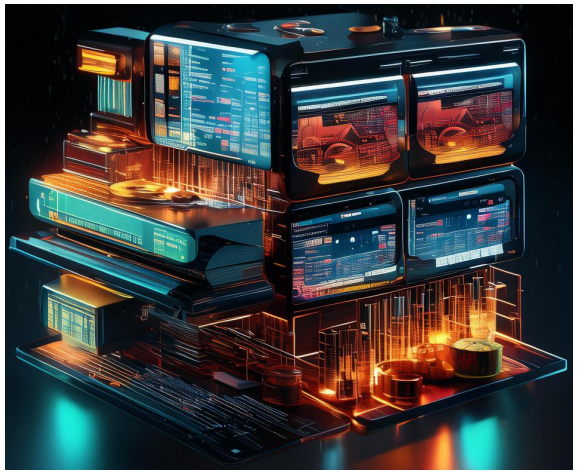
Multi-modal Foundation Model

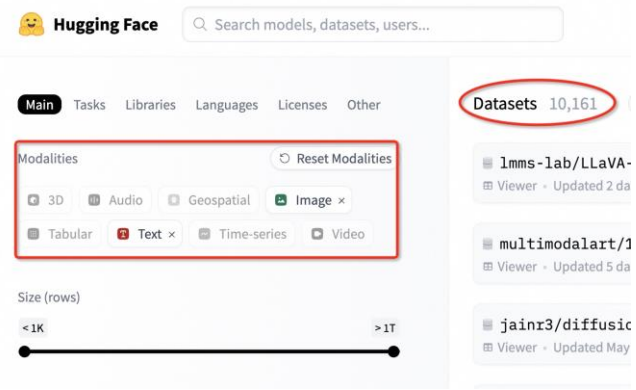


Model-centric → Data-centric → Data-Model Co-development

[1] (arXiv:2407.08583) The Synergy between Data and Multi-Modal Large Language Models: A Survey from Co-Development Perspective

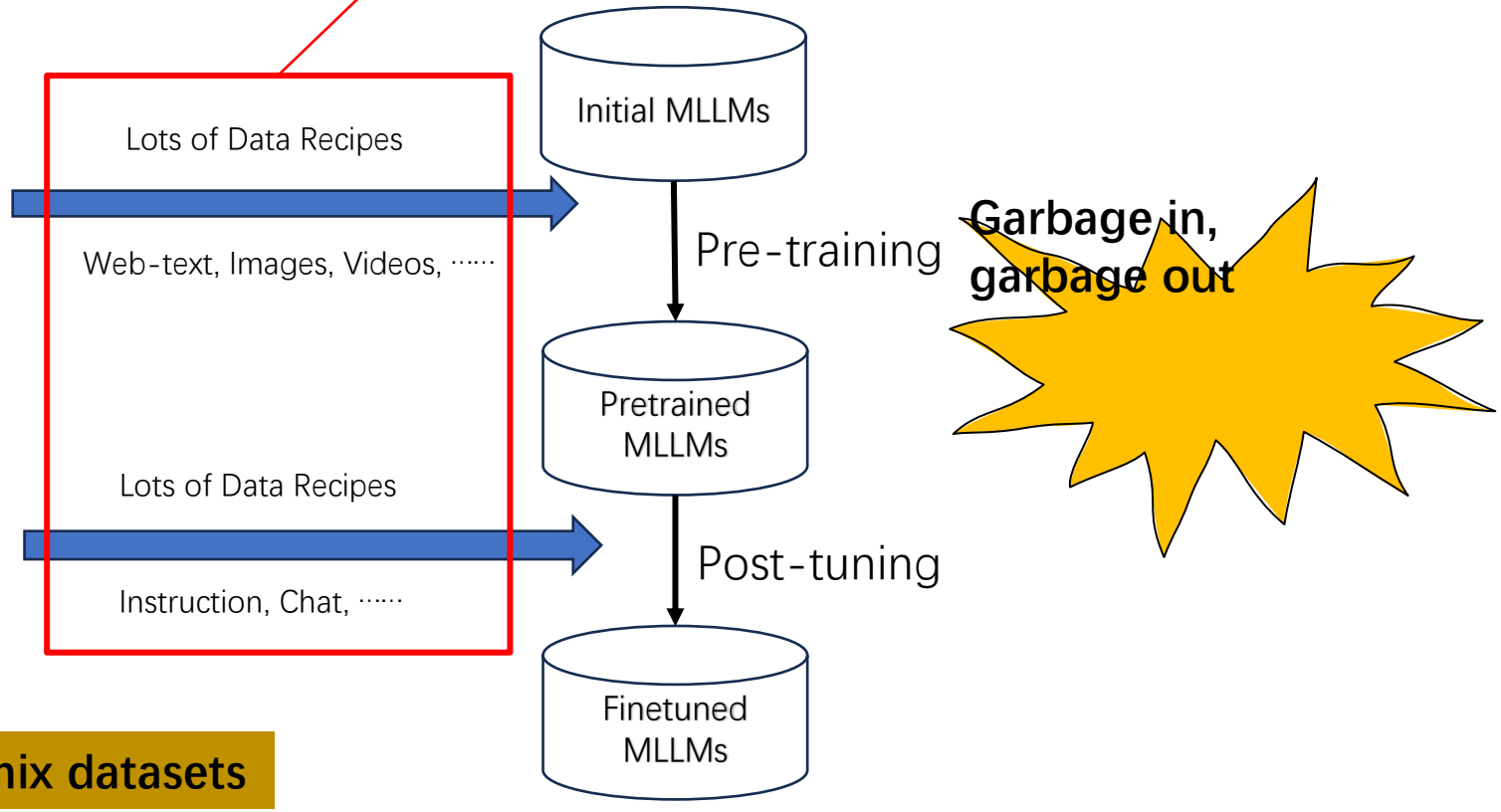
From Data to MLLMs





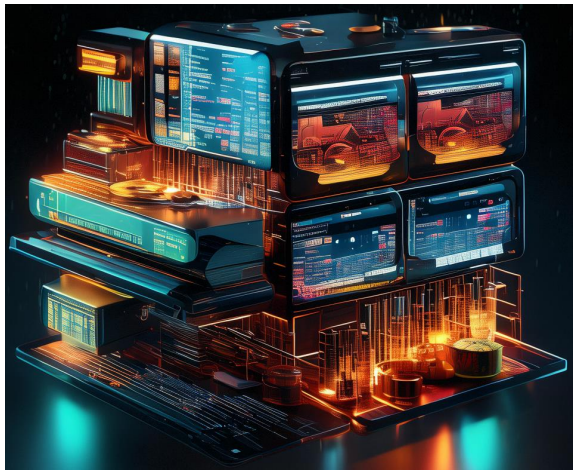
Data Recipe: process & mix datasets

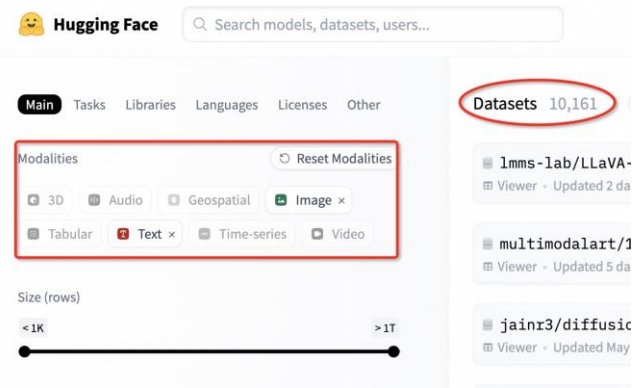
- **Different types** of data from **diverse** sources
- **High heterogeneity** in data recipes



[2] (SIGMOD'24) Data-juicer: A one-stop data processing system for large language models
[3] (arXiv:2405.14908) Data Mixing Made Efficient: A Bivariate Scaling Law for Language Model Pretraining

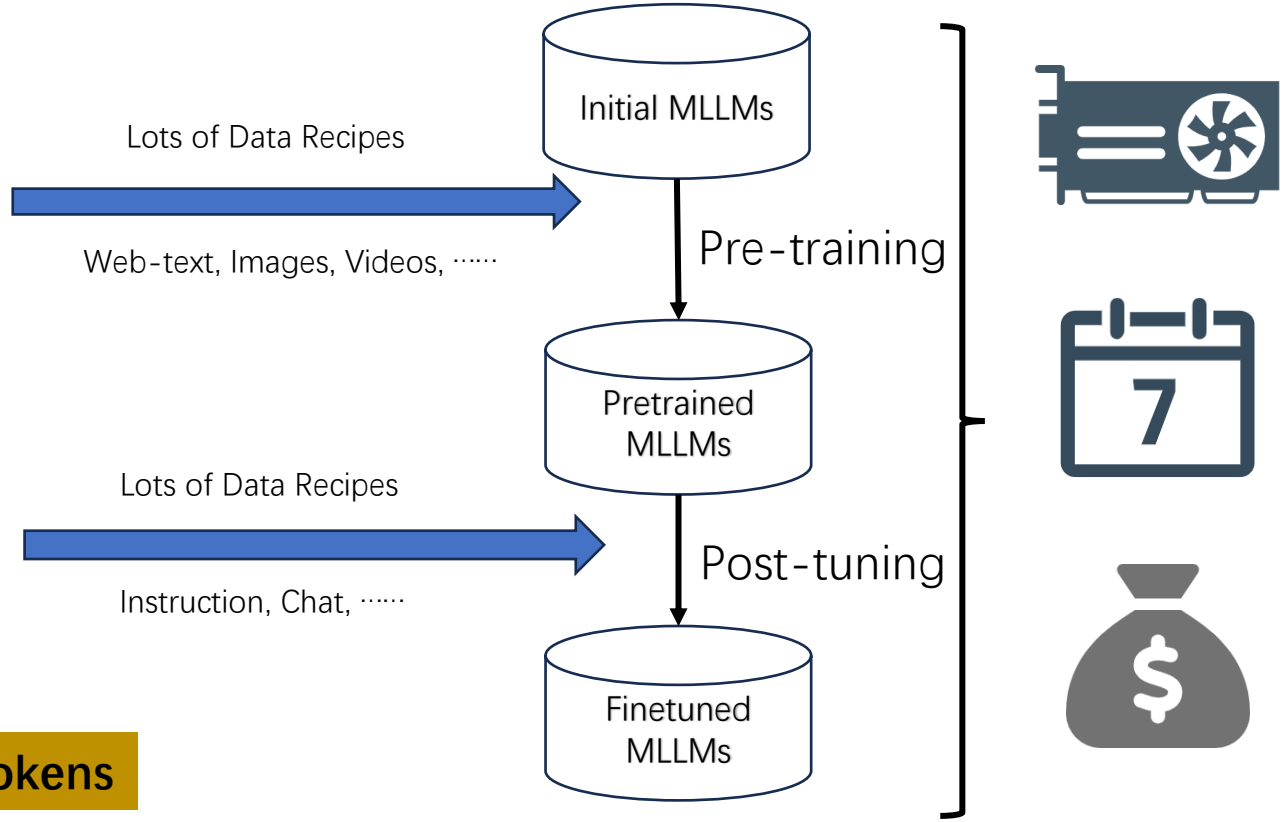
From Data to MLLMs





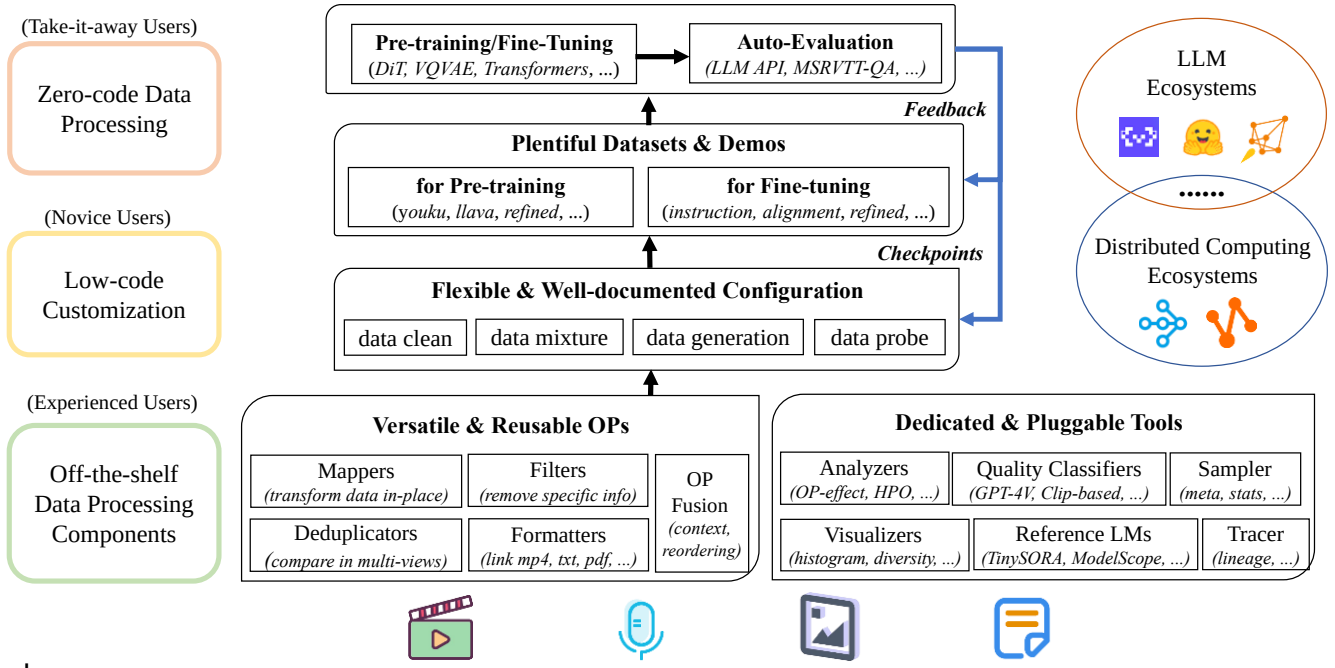
Billion/Trillion Tokens

- **Massive data volume**
- **High cost** to retry and refine



The Data-Juicer System

- **Systematic & Reusable**
  - 100+ standard operators
  - 50+ tools & data recipes
  - Data analysis, cleaning, synthesis
- **Efficient & Actionable**
  - Industrial performance optimization
  - Distributed data processing
  - Data-model co-development
- **Effect-Proven & Instructional**
  - Rank-1 text-to-video model on VBench
  - Beats GPT4v/Gemini on MMVP bench
  - Maintained survey & practical insights



[2] (SIGMOD'24) **Data-juicer: A one-stop data processing system for large language models**

Tutorial Outline

- **Data Processing Foundations**
  - Building Blocks of Data Processing: Data-Juicer's Operators
  - Composing Atomic Capabilities: Data-Juicer's Data Recipes
- **Advanced Data Processing**
  - Exploring Data Recipes: The Data-Juicer Sandbox Lab
  - From Exploration to Production: High-Performance Data Factory
- **Use Cases:** From Text to Video Data Processing
- **Resources and Conclusion**

Jupyter notebooks embedded

Tutorial Outline

- **Data Processing Foundations**
  - **Building Blocks of Data Processing: Data-Juicer's Operators**
  - Composing Atomic Capabilities: Data-Juicer's Data Recipes

- **Advanced Data Processing**
  - Exploring Data Recipes: The Data-Juicer Sandbox Lab
  - From Exploration to Production: High-Performance Data Factory

- **Use Cases:** From Text to Video Data Processing

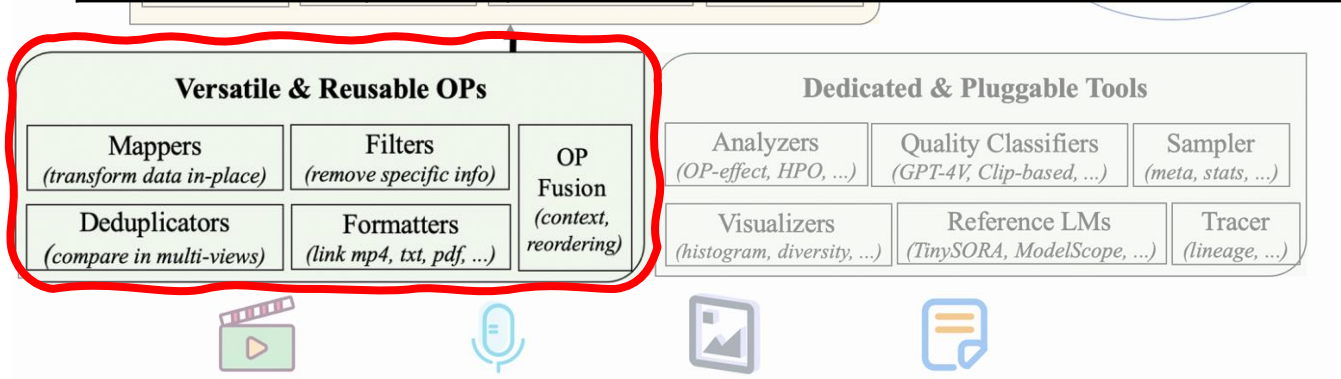
- **Resources and Conclusion**

Data-Juicer: Standard OPs

- **Standard** Interface
- Basic **capabilities** of data processing

- Input: datasets
- Output: processed datasets

Type	Num.	Description
Formatter	7	Discovers, loads, canonicalizes source data
Mapper	46	Edits and transforms samples
Filter	43	Filters out low-quality samples
Deduplicator	5	Detects and removes duplicate samples
Selector	4	Selects top samples based on ranking



From Formatters to Dataset

- Unified format, path-based storage
- Text as anchor, expressing relationships
  - Alignment & Position
  - Local & Global
- Built-in typical dataset conversion tools
  - Video-ChatGPT <--> Data-Juicer
  - Youku-CN <--> Data-Juicer
  - InternVid <--> Data-Juicer
  - LLaVA <--> Data-Juicer
  - ……


```
{
  "text": "<vid> 欢迎来到杭州！ <|eoc|> <vid> 这是美丽的杭州西湖，壮观的钱塘江大潮。 <|eoc|> 让我们目光聚焦到市区西北部。 <|eoc|> <aud> 聚焦到阿里云！<vid> <|eoc|> <img> 这里则是坐落于杭州市西湖区的阿里巴巴云谷园区。 <|eoc|>",
  "videos": [
    "path/to/the/video/of/hangzhou",
    "path/to/the/video/of/WestLake-and-QianTang-River",
    "path/to/the/video/of/CloudValley"
  ],
  "audios": [
    "path/to/the/audio/of/welcome-to-alibaba-cloud"
  ],
  "images": [
    "path/to/the/image/of/AlibabaCloud-logo",
  ],


  "meta": {
    "src": "customized",
    "version": "0.1",
    "author": "xxx"
  },
  "stats": {
    "lang": "zh",
    "lang_score": 0.9645169377,
    "video_frames_text_matching_score": 0.8,
    "video_duration": 15.5,
    "avg_audio_image_match_score": 0.7,
    "image_aesthetics_scores": 0.6
  }
}
```

Data-Juicer: Standard OPs

- Standard OP base class

```python
class OP:
    ...
    def __call__(self, dataset: DJDataset) -> DJDataset:
        ...
        return self.run(dataset, **args)


class Mapper(OP):
    ...
    def process(self, sample: Dict) -> Dict:
        # process a single sample
        ...

    def run(self, dataset: DJDataset) -> DJDataset:
        return dataset.map(self.process, **args)
    ...

class Filter(OP):
    ...
    def compute_stats(self, sample: Dict) -> Dict:
        # compute stats of a single sample
        ...

    def process(self, sample: Dict) -> bool:
        # decide whether to keep this single sample
        ...

    def run(self, dataset: DJDataset) -> DJDataset:
        return dataset.map(self.compute_stats, **args)
                      .filter(self.process, **args)
    ...
```

- Flexible invoking logic

```python
op1_config = ...
op2_config = ...

dataset = Dataset.from_jsonl(path=..., mode="standalone/ray")

op1 = Operator_1(**op1_config)
op2 = Operator_2(**op2_config)

# single OP
dataset = dataset.process(op1)
dataset = dataset.process(op2)

# multiple OPs
dataset = dataset.process(op1).process(op2)
# or
dataset = dataset.process([op1, op2])

dataset.export(path=..., format="jsonl")
```

Data Processing Hands-on

- Jupyter notebooks part 1
  - 1.1 OP insights about specific OP types, and multi-modal OP examples
  - 1.2 datasets loading: multi-modal dataset format <--> Data-Juicer format
  - 1.3 Brief introduction of how to develop a new OP

Tutorial Outline

- **Foundational Abilities**
  - Building Blocks of Data Processing: Data-Juicer's Operators
  - **Composing Atomic Capabilities: Data-Juicer's Data Recipes**

- **Advanced Data Processing**
  - Exploring Data Recipes: The Data-Juicer Sandbox Lab
  - From Exploration to Production: High-Performance Data Factory

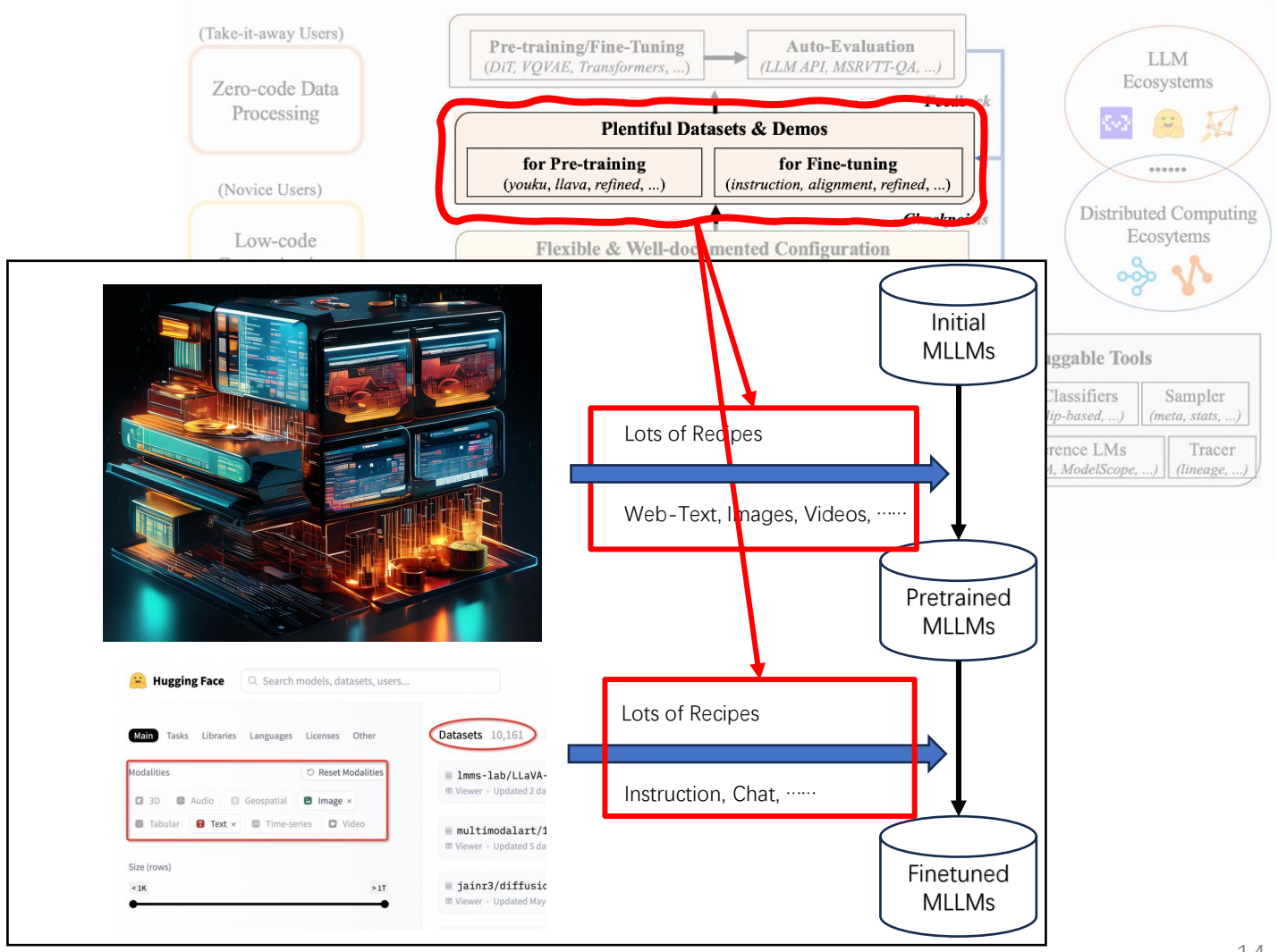
- **Use Cases:** From Text to Video Data Processing

- **Resources and Conclusion**

Data-Juicer: Data Recipes

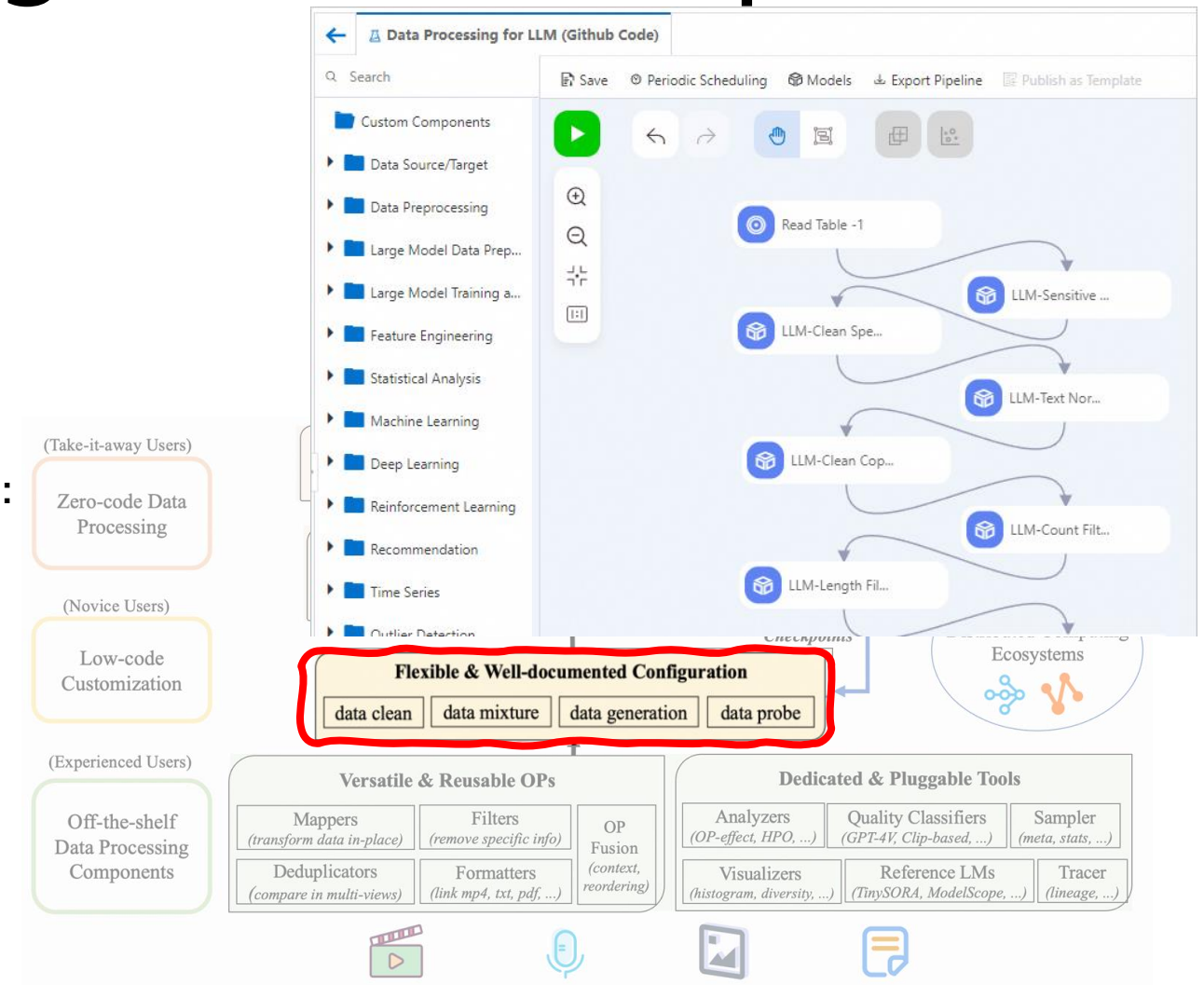
- Data recipe includes:
  - Data source
  - Mixture ratios
  - Data processing pipeline
  - ……

- **Effect-proven**
- **Reusable** and **shared** for the community



How to config for data recipes

- Three ways to config:
  - A piece of code: describe the processing pipeline
  - A config file (**recommended: simple & flexible**)
  - Drag-drop interface: equivalent to a config file, (available on PAI-Designer)



Example: A Config File for Data Recipe

```yaml
project_name: 'video-dj-recipe-demo'
dataset_path: 'video-dj-recipe-youku.jsonl'
export_path: 'video-dj-recipe-youku-generated.jsonl'
np: 42                                  # number of subprocess to process your data
open_tracer: true
video_key: 'videos'                     # Key name of field to store the list of sa
video_special_token: '<video>'          # The special token that represents an vide
eoc_special_token: '<|__dj__eoc|>'      # The special token that represents the end

# process schedule: a list of several process operators with their arguments
process:
  - video_deduplicator:
  - fix_unicode_mapper:                 # fix unicode errors in text.
  - punctuation_normalization_mapper:   # normalize unicode punctuations to English

  - video_split_by_scene_mapper:        # split videos into scene clips
      detector: 'ContentDetector'       # PySceneDetect scene detector. Should be one o
      threshold: 27.0                   # threshold passed to the detector
      min_scene_len: 15                 # minimum length of any scene
      show_progress: false              # whether to show progress from scenedetect
  - video_captioning_from_audio_mapper: # caption a video according to its audio streams
      keep_original_sample: true        # whether to keep the original sample. If it's
  - video_tagging_from_audio_mapper:    # Mapper to generate video tags from audio stream
      hf_ast: 'MIT/ast-finetuned-audioset-10-10-0.4593'  # Huggingface model name for the audio classifi
  - video_frames_text_similarity_filter: # keep samples those similarities between sampled
      hf_clip: 'openai/clip-vit-base-patch32'  # clip model name on huggingface to compute the
      min_score: 0.1                    # the min similarity to keep samples.
      max_score: 1.0                    # the max similarity to keep samples.
      frame_sampling_method: 'all_keyframes'  # sampling method of extracting frame images fr
      frame_num: 3                      # the number of frames to be extracted uniforml
      horizontal_flip: false            # flip frame image horizontally (left to right)
```

Global Arguments:
- Paths to: input & output datasets, ……
- Num of workers: process in parallel
- ……

Operators in Ordered List:
- OP name
- Arguments of OPs
- Including several categories of OPs

Filtering (single-modality quality):
- video de-dup; text de-noise; ……

Enhancing (diversity)
- video clips; re-captioning; ……

Filtering (cross-modality quality):
- text-video matching degree; ……

Built-in Data Recipes

dataset	#samples before	#samples after	keep ratio	data recipe link	data link	source
arXiv	1,724,497	1,655,259	95.99%	redpajama-arxiv-refine.yaml	Aliyun ModelScope HuggingFace	Redpajama
····· (Textual Data Recipes)						
MGM pretrain (1.2M)	1,266,268	159,288 (repeat 4 times)	12.58%	mgm-pretrain-top2-ops-refine.yaml	Aliyun ModelScope HuggingFace	LLaVA-1.5 ALLaVA-4V
Video-text	1,217,346	147,176	12.09%	2_multi_op_pipline.yaml	Aliyun ModelScope HuggingFace	Panda-70M InternVid MSR-VTT
····· (Multi-modal Data Recipes)						

- Refined for better diversity
  - Combination of SOTA dataset, e.g., Panda-70M, InternVid, MSR-VTT, ...
  - Contrastive data synthesis
  - .....
- Refined for higher quality
  - Data cleaning guided by OPs, tools, sandbox, ...
  - Duplicating good data
  - .....

Effect-Proven Data Recipes

dataset	#samples before	#samples after	keep ratio	data recipe link	data link	source
arXiv	1,724,497	1,655,259	95.99%	redpajama-arxiv-refine.yaml	Aliyun ModelScope HuggingFace	Redpajama
····· (Textual Data Recipes)						
MGM pretrain (1.2M)	1,266,268	159,288 (repeat 4 times)	12.58%	mgm-pretrain-top2-ops-refine.yaml	Aliyun ModelScope HuggingFace	LLaVA-1.5 ALLaVA-4V
Video-text	1,217,346	147,176	12.09%	2_multi_op_pipline.yaml	Aliyun ModelScope HuggingFace	Panda-70M InternVid MSR-VTT
····· (Multi-modal Data Recipes)						

- Video-text recipe example[4]
  - Rank-1 model (ours) on Vbench leaderboard
- Image-text recipe example[5]
  - 50.7 (ours) v.s. 38.7 (GPT-4V) on MMVP benchmark

[4] (arXiv:2407.11784) Data-Juicer Sandbox: A Comprehensive Suite for Multimodal Data-Model Co-development
[5] (arXiv:2408.04594) ImgDiff: Contrastive Data Synthesis for Vision Large Language Models

Data Processing Hands-on

- Jupyter notebooks part 2
  - 2.1 conduct data processing with existing recipes
  - 2.2 introduction of some basic global information
  - 2.3 how to modify your own recipes

Tutorial Outline

- **Foundational Abilities**
  - Building Blocks of Data Processing: Data-Juicer's Operators
  - Composing Atomic Capabilities: Data-Juicer's Data Recipes

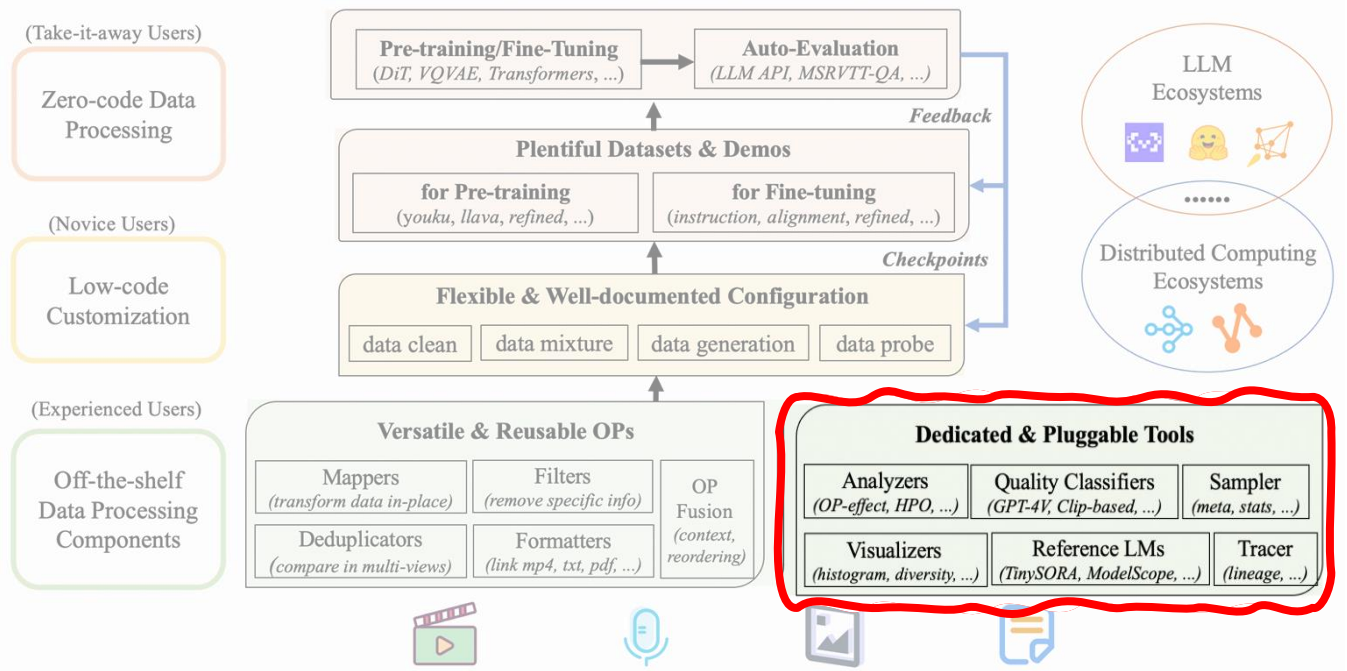
- **Advanced Data Processing**
  - **Exploring Data Recipes: The Data-Juicer Sandbox Lab**
  - From Exploration to Production: High-Performance Data Factory

- **Use Cases:** From Text to Video Data Processing

- **Resources and Conclusion**

Data-Juicer: Tools

- **Dedicated** for **complex** and **specific** functions, supporting data recipe exploration
  - Analysis
  - Processing trace
  - Visualization
  - ……



Tools: Tracer

- Sample-level changes before/after each OP

Filter: filtered-out samples


```
{
    "text":"你好，请问你是谁",
    "meta":{
        "src":"customized",
        "date":null,
        "version":null,
        "author":"xxx"
    },
    "stats":{
        "lang":"zh",
        "lang_score":0.9450979829
    }
}
```


Mapper: modified samples

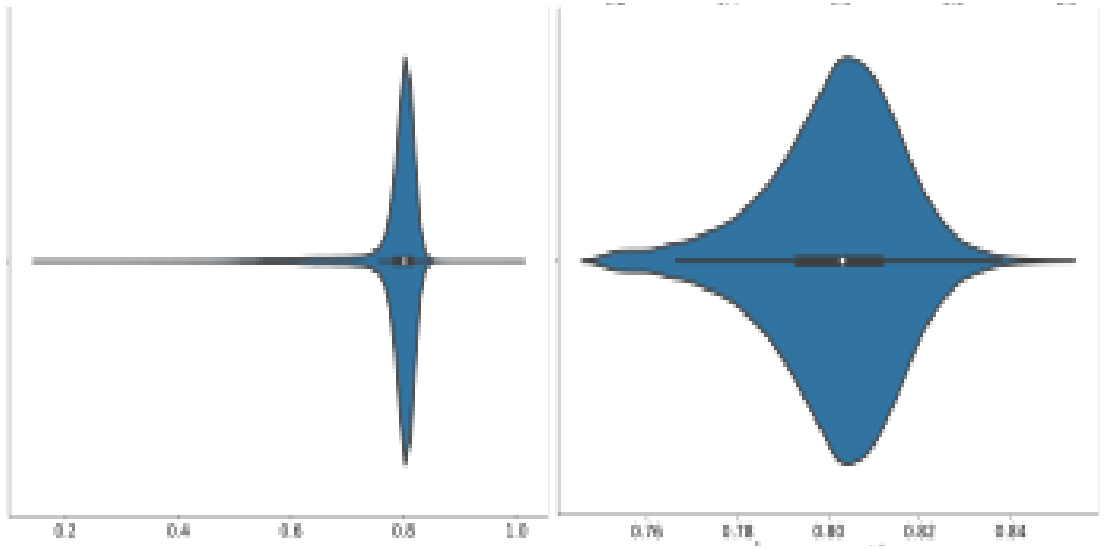

```
{
    "original text":"This is the email of our company: euqdh@cjqi.com",
    "processed_text":"This is the email of our company: "
}
```


Deduplicator: (near-) duplicate sample pairs

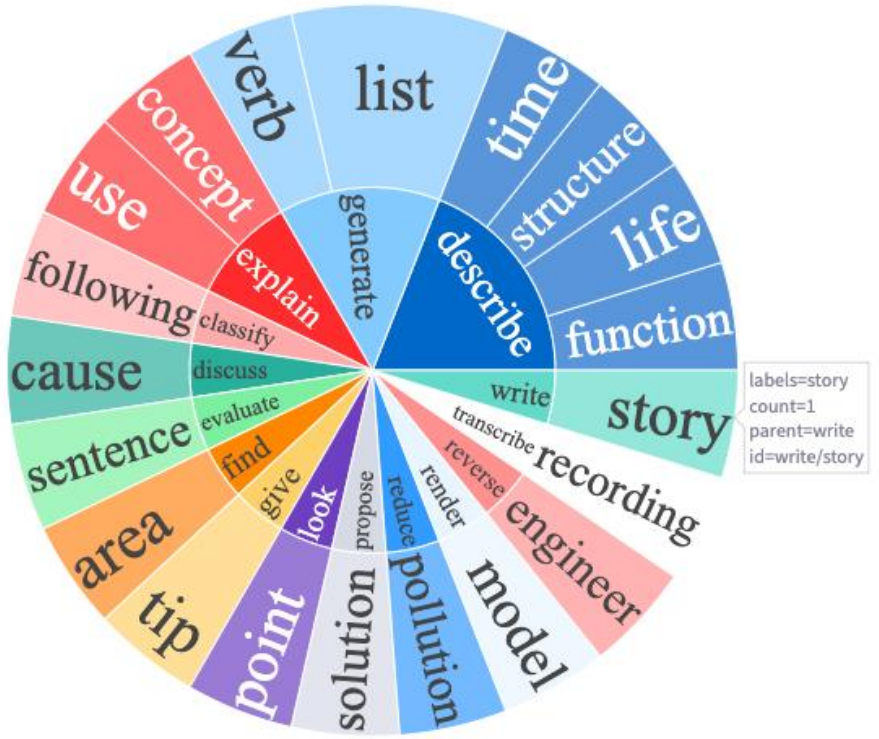

```
{
    "dup1":{
        "text":"This paper proposed a novel method on LLM pretraining.",
        "meta":{
            "meta_inner":{
                "src":"customized",
                "date":null,
                "version":null,
                "author":"xxx"
            }
        },
        "stats":{
            "lang":"en",
            "lang_score":0.9568607211
        },
        "hash":"df544ffbc314a6d27b2847429246be76"
    },
    "dup2":{
        "text":"This paper proposed a novel method on LLM pretraining.",
        "meta":{
            "meta_inner":{
                "src":"customized",
                "date":null,
                "version":null,
                "author":"xxx"
            }
        },
        "stats":{
            "lang":"en",
            "lang_score":0.9568607211
        },
        "hash":"df544ffbc314a6d27b2847429246be76"
    }
}
```

Tools: Analyzer

- **Quality**: statistics distribution
  - Character repetition ratio
  - Perplexity
  - ……



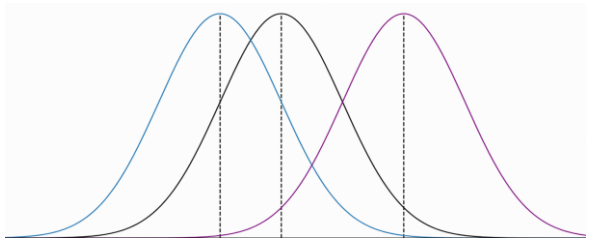
- **Diversity**: lexical combination



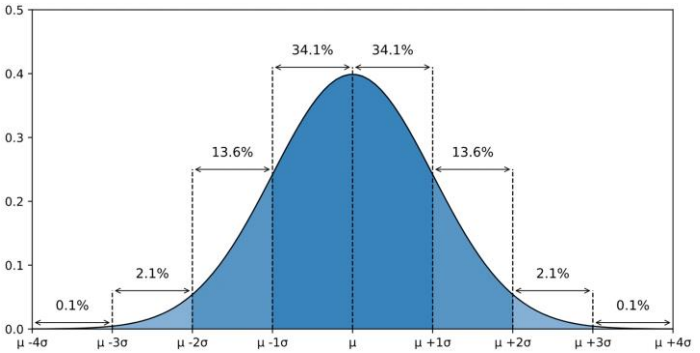
Tools: Recipe Iteration

K-Sigma & Model Sensitivity

- **Analyze** data automatically
  - Compute quantitative values of each feature dimension
  - Statistics (fluency, aesthetic, cross-modal matching, etc.)
  - Sample data and evaluate the sensitivities from models
- **Filter out** data outside the k-sigma range



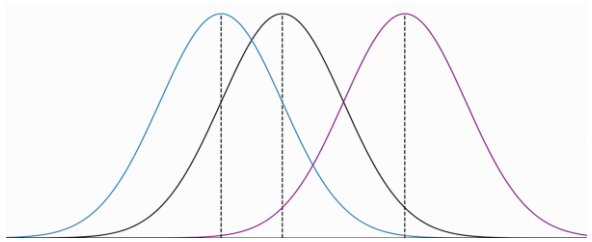
stats from 40+ Data-Juicer Filters



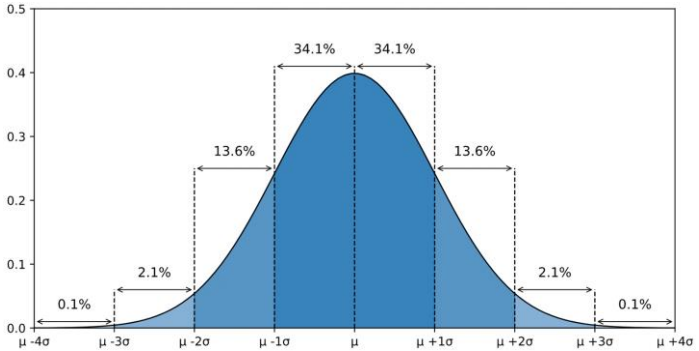
Tools: Recipe Iteration

K-Sigma & Model Sensitivity

- **Analyze** data automatically
  - Compute quantitative values of each feature dimension
  - Statistics (fluency, aesthetic, cross-modal matching, etc.)
  - Sample data and evaluate the sensitivities from models
- **Filter out** data outside the k-sigma range

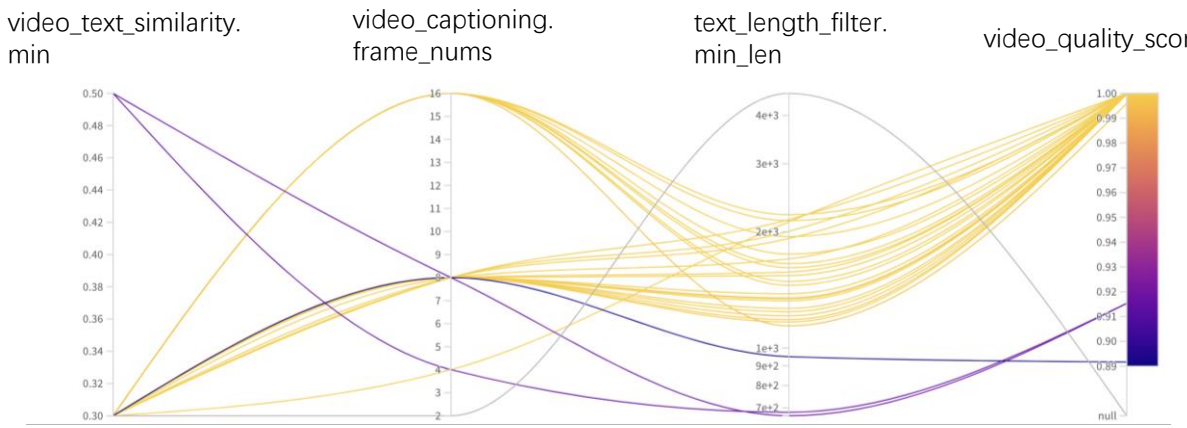


stats from 40+ Data-Juicer Filters



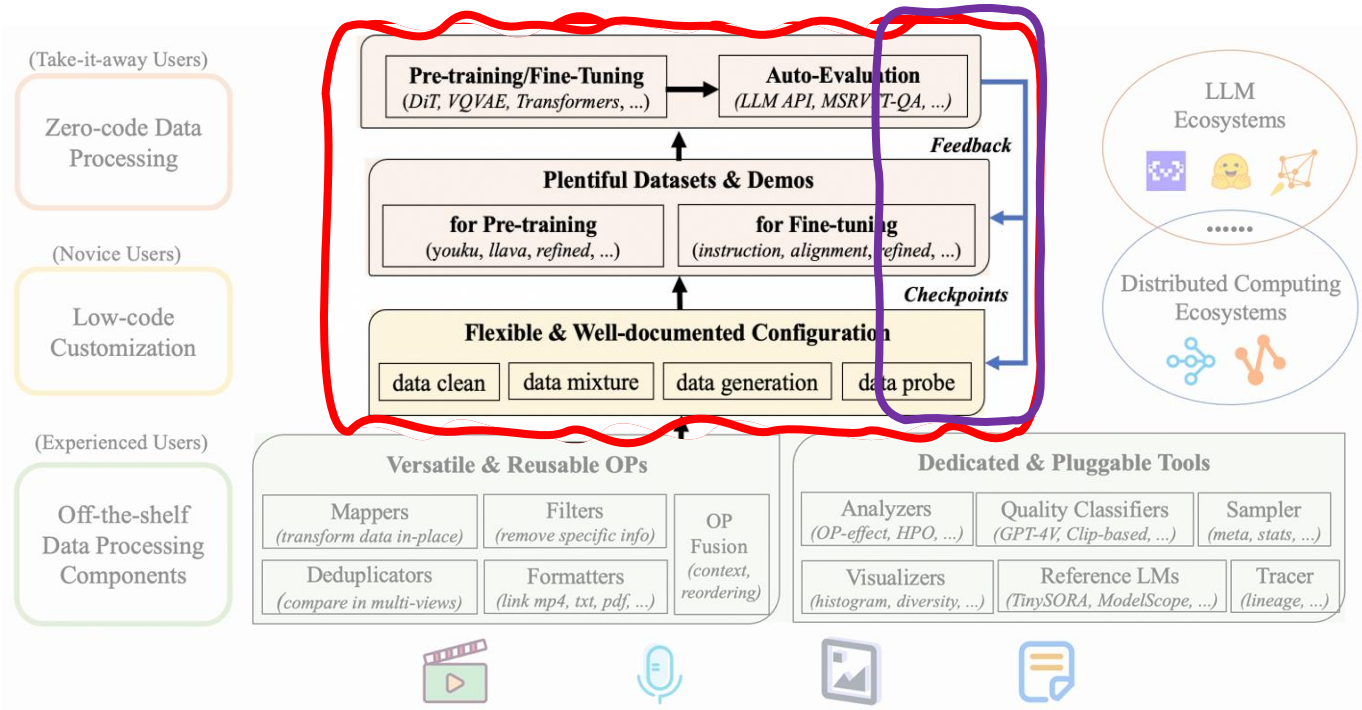
Auto Recipe Optimization

- **Analyze** importance & relevance of hyper-parameters in data recipe
- Integrated W&B Sweeps: support heuristic search & Bayesian opt.



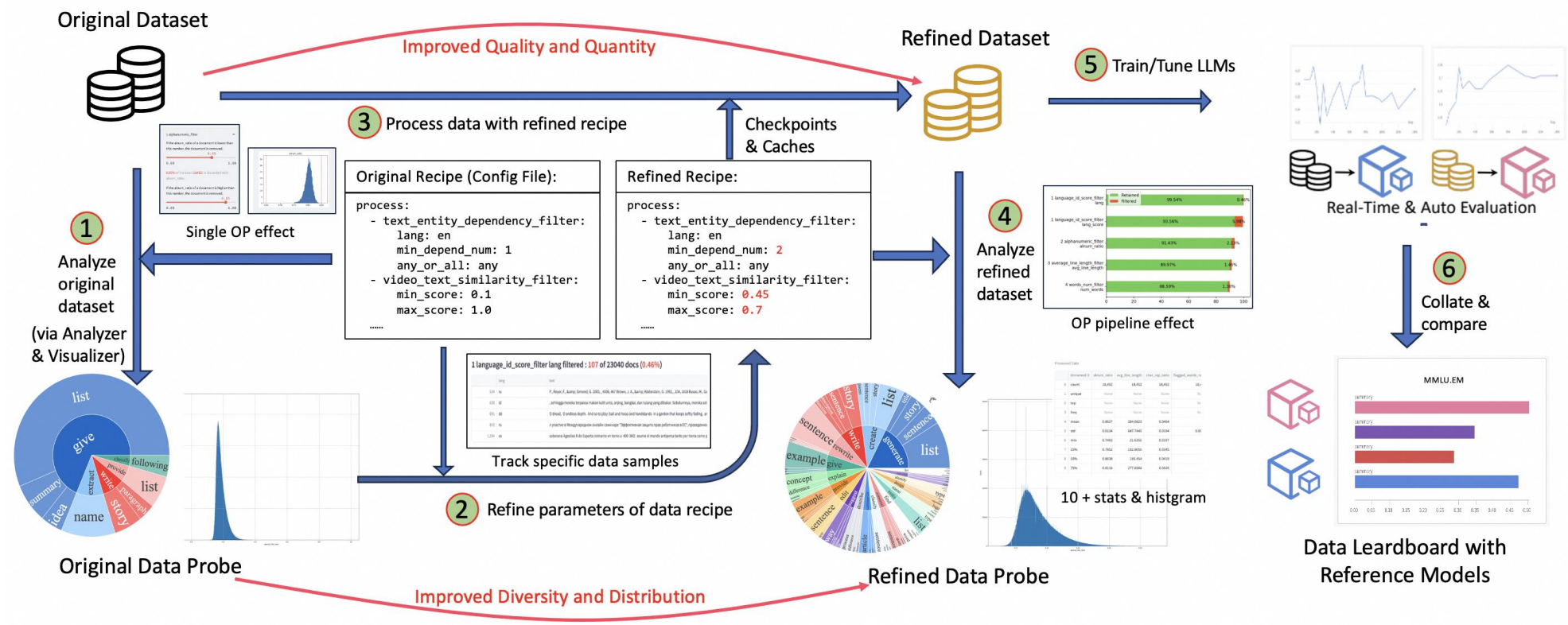
Data-Juicer: Feedback Loop

- Return multiple **feedback signals**
- Refine recipe **systematically**



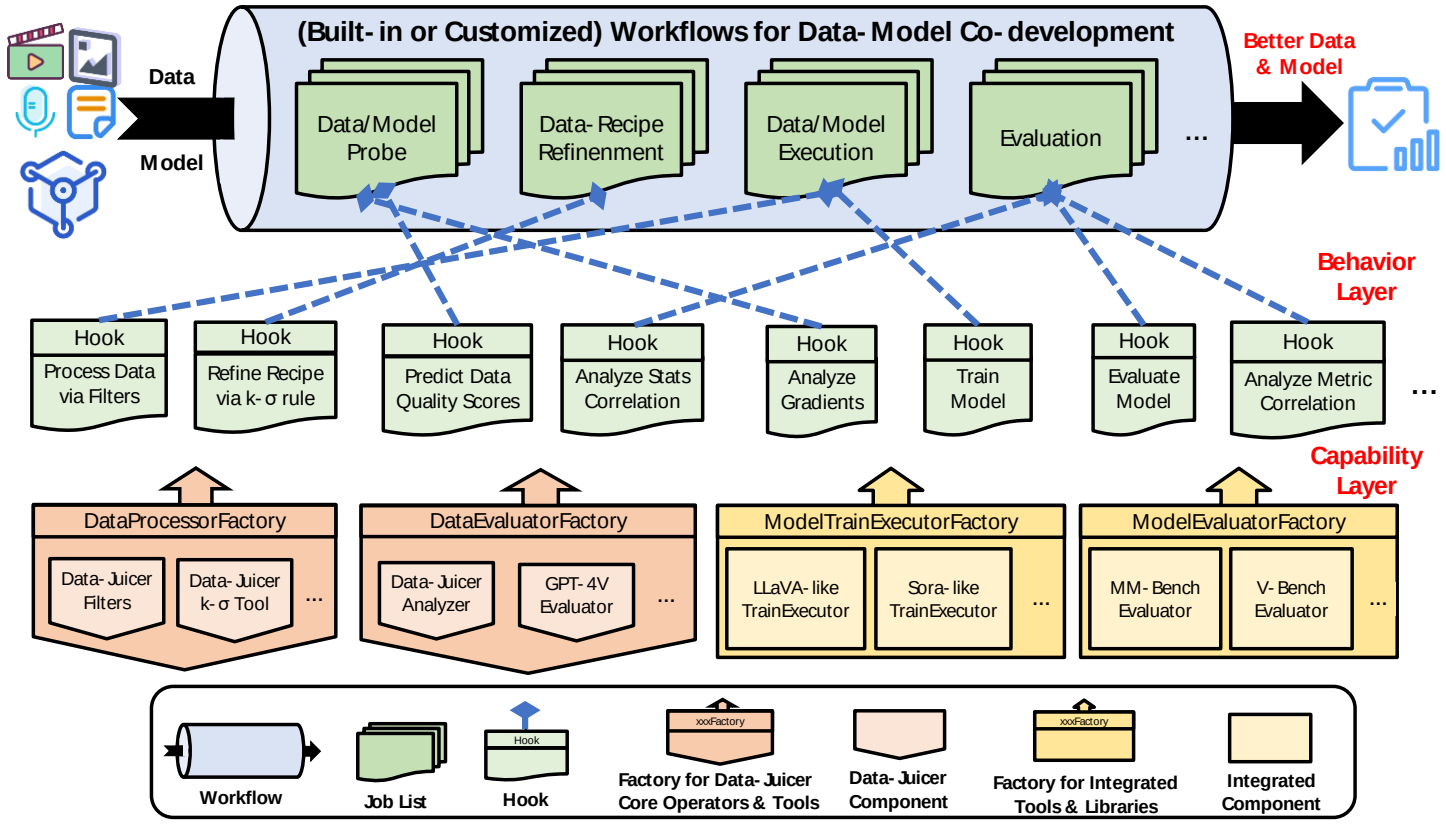
Data-Juicer: Feedback Loop

- Improve data quality with feedbacks from different steps of data-model co-development
- The whole loop has been formed as a **Data-Model Sandbox** in Data-Juicer now!



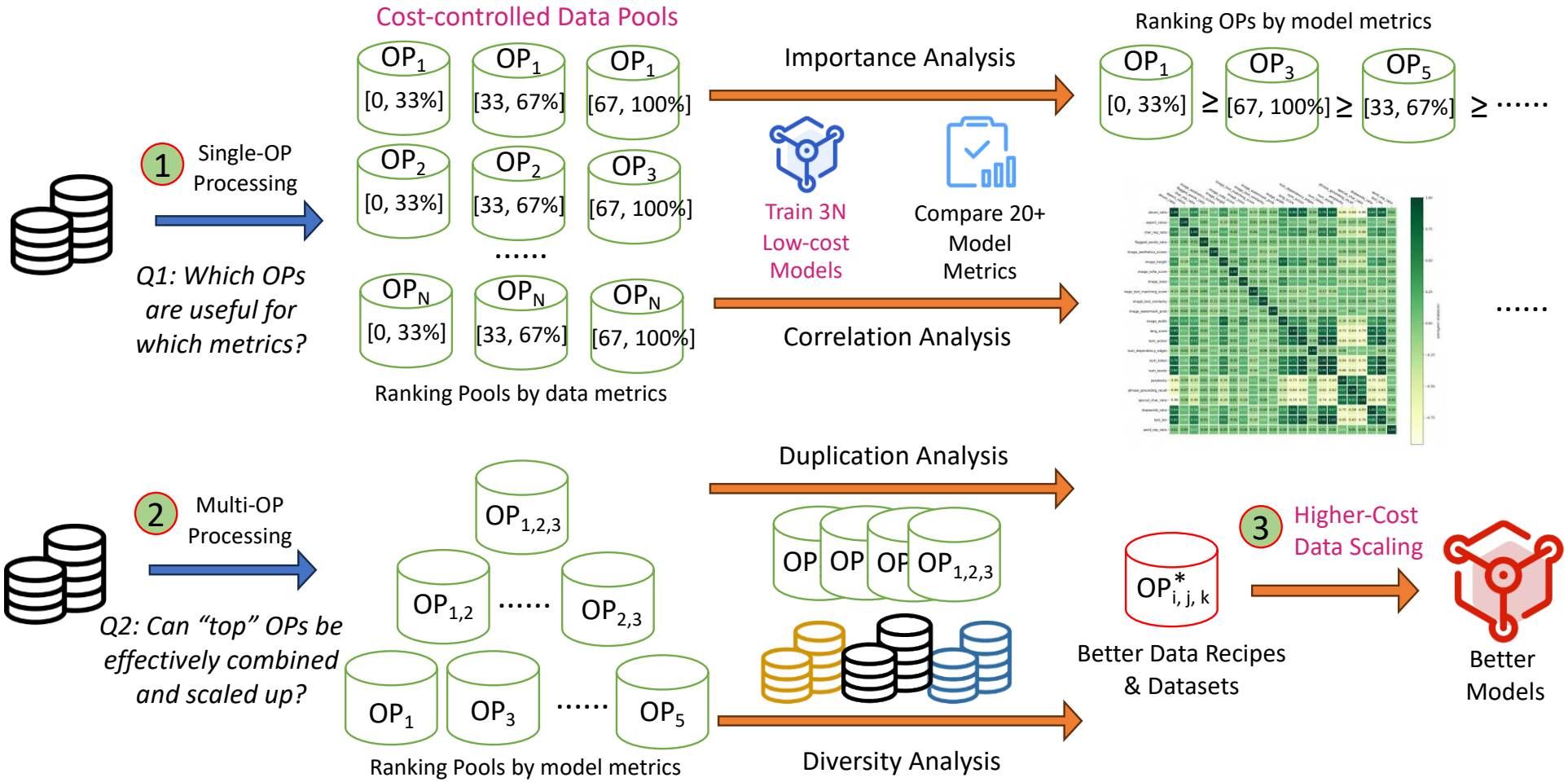
Data-Juicer: Sandbox Lab

- Three-layer Orchestration
  - Workflow
  - Behavior
  - Capability
- Cost-controlled
- Easy Customization



[3] (arXiv:2407.11784) Data-Juicer Sandbox: A Comprehensive Suite for Multimodal Data-Model Co-development

Data-Juicer: Sandbox Lab



Recipe Exploration Hands-on

- Jupyter notebooks part 3
  - 3.1 how to use the Data-Juicer tools
  - 3.2 how to use the Data-Juicer SandBox

Tutorial Outline

- **Foundational Abilities**
  - Building Blocks of Data Processing: Data-Juicer's Operators
  - Composing Atomic Capabilities: Data-Juicer's Data Recipes

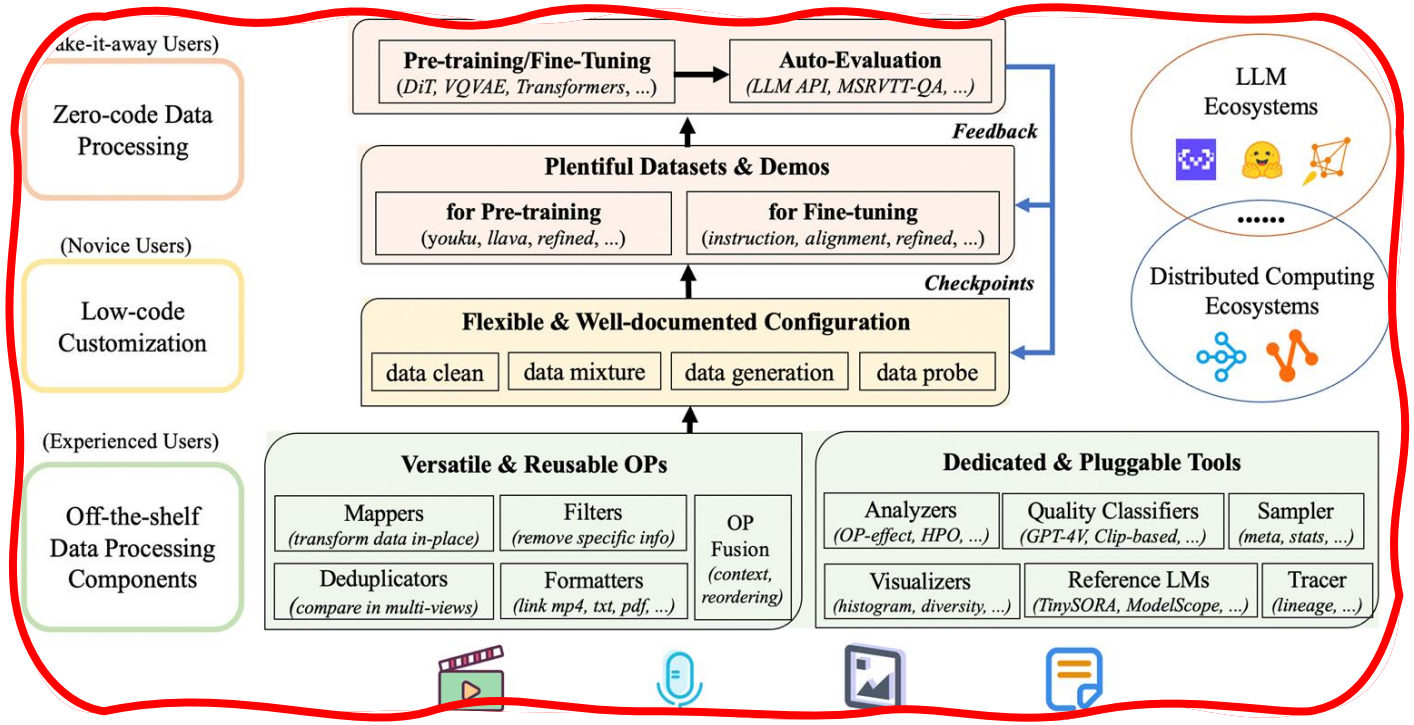
- **Advanced Data Processing**
  - Exploring Data Recipes: The Data-Juicer Sandbox Lab
  - **From Exploration to Production: High-Performance Data Factory**

- **Use Cases:** From Text to Video Data Processing

- **Resources and Conclusion**

Data-Juicer: Processing Optimization

- **Distributed Processing**
  - Systematic scalability
- **OP Fusion**
  - Avoid redundant computation
- **Resource Management**
  - Adaptively assign resources
- **Fault Tolerance**
  - Increase reliability

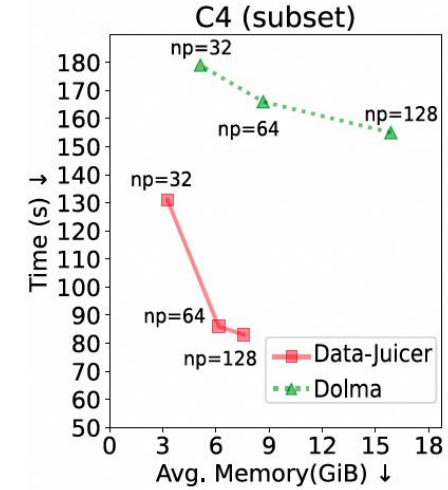


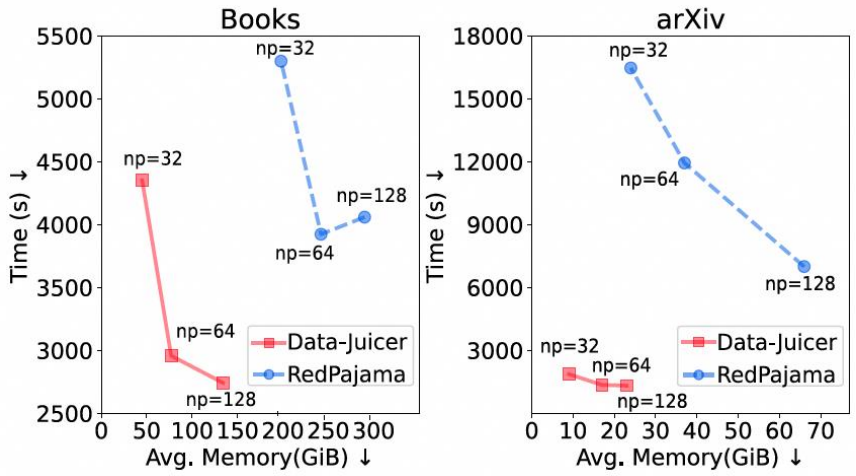
Towards a Production System

Processing Performance

Performance Comparison

- Standalone mode, optimization flags off
- Data Processing Baseline: Dolma & RedPajama
  - Avg. Time (↓ 50.6%)
  - Avg. Memory (↓ 55.1%)
- The basic implementation is already efficient!





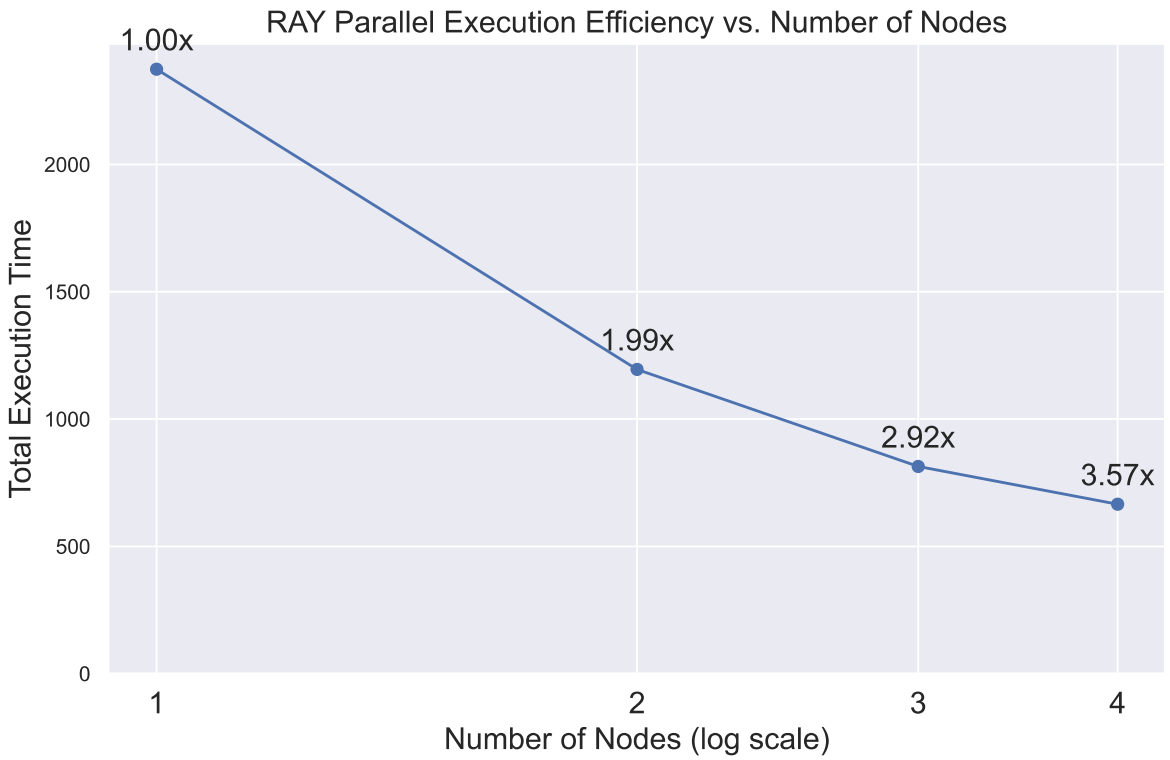
Optimization: Parallelization

Opt 1: Parallel Processing

- Multi-core multi-node
- Support CUDA

- Systematic scalability with Ray

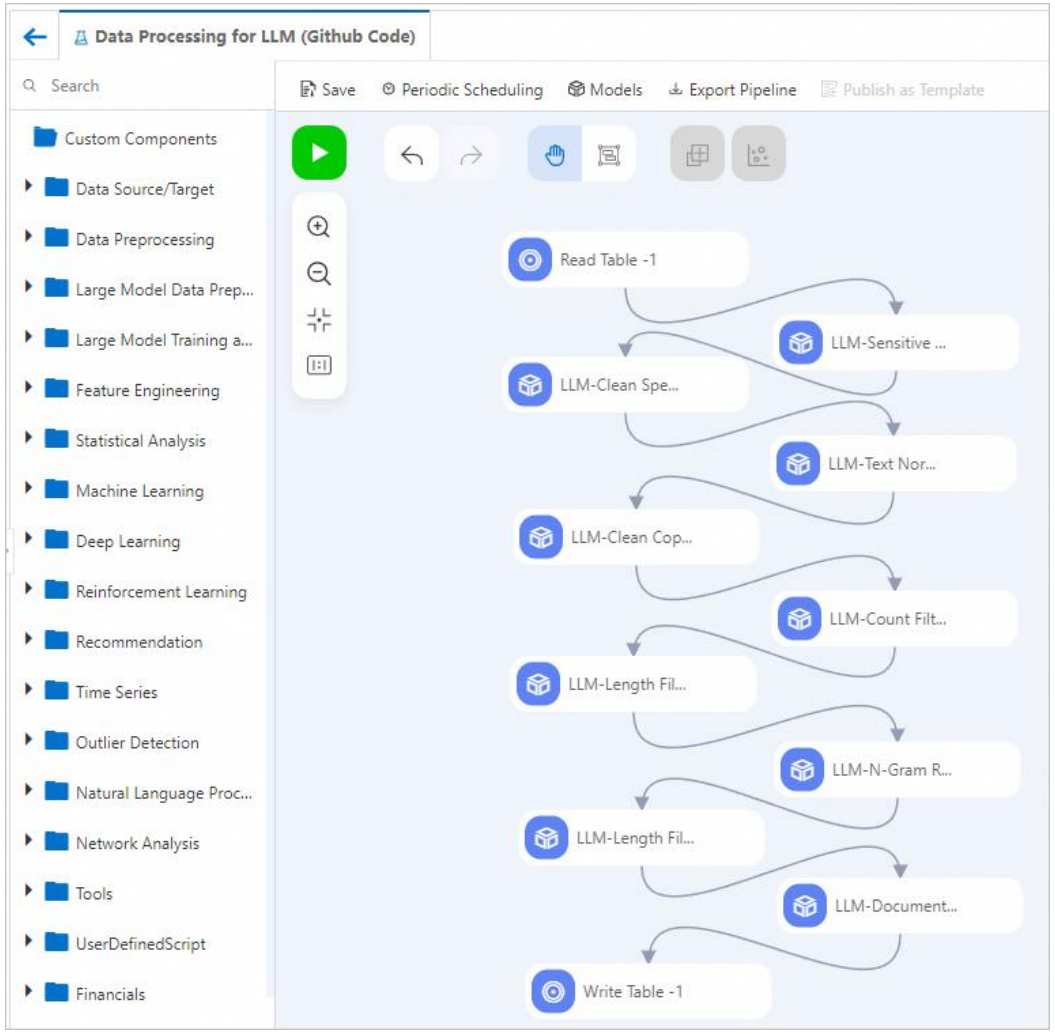
  (https://github.com/ray-project/ray)



Optimization: Parallelization

Opt 1: Parallel Processing

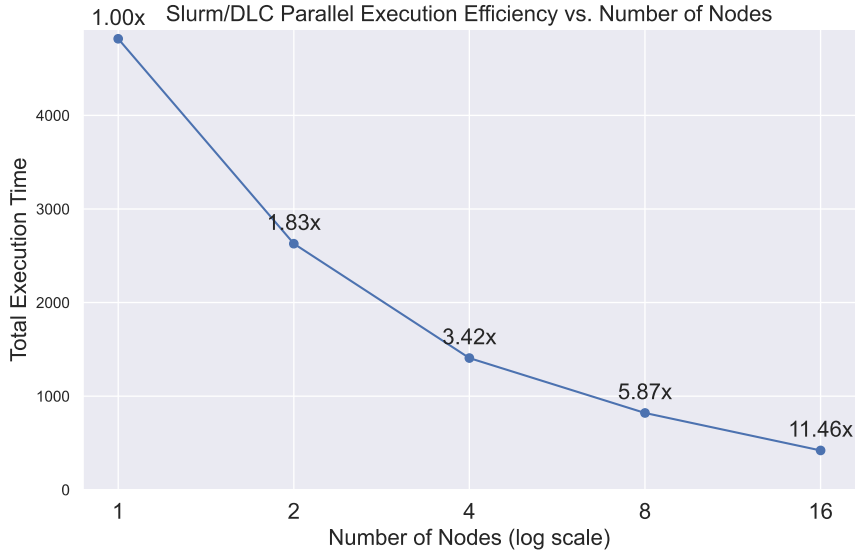
- On-cloud distributed processing:
  - Alibaba Cloud PAI-Designer

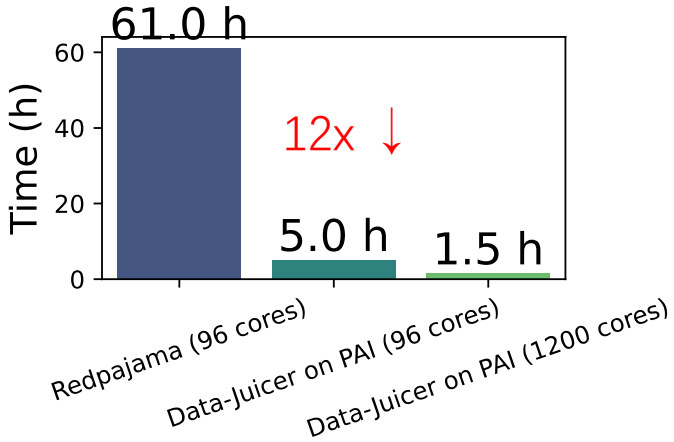


Optimization: Parallelization

Opt 1: Parallel Processing

- On-cloud distributed processing:
  - Alibaba Cloud PAI-Designer
  - Alibaba Cloud PAI-DLC/DSW

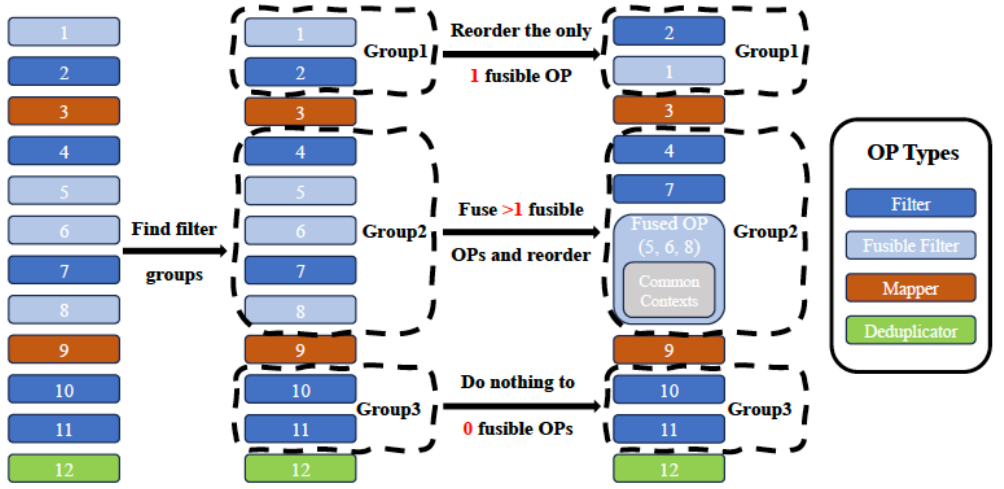




Optimization: OP Fusion

Opt 2: OP Fusion + Context Management + OP Reordering

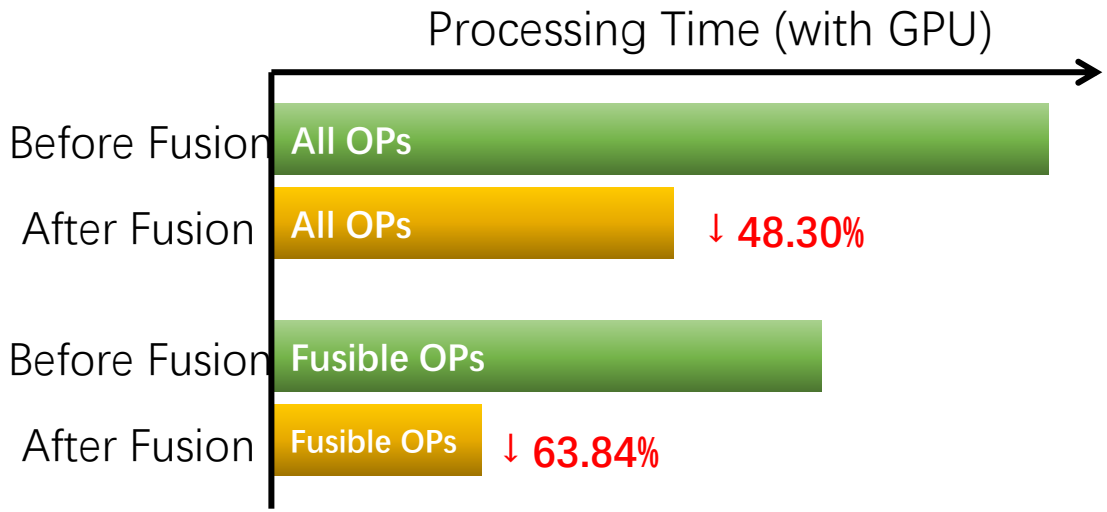
- Core facts:
  - Some OPs share the same contexts & computation sub-procedures
  - Filters are commutative
  - Some OPs are much more time-consuming



Optimization: OP Fusion

Opt 2: OP Fusion + Context Management + OP Reordering

- End-to-end Performance Comparsion
- MSR-VTT, 7 OPs (5 capable of fusion)
  - Video Frame-Text Similarity Filter
  - Video NSFW Filter
  - Video Motion Score Filter
  - Video OCR Area Ratio Filter
  - Video Watermark Filter



Optimization: Resource Management

Opt 3: OP-aware parallelization & Quantization

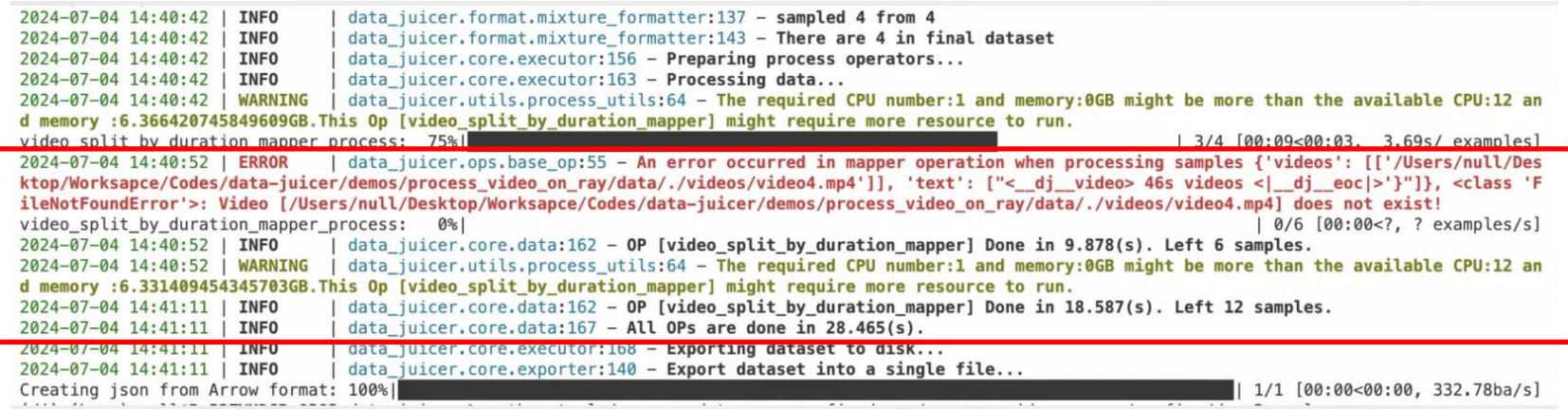
- OP-aware Parallelization
  - different OP requires different amount of resources
  - adaptively adjust parallelization strategy for different OPs

- Low Precision Supported → Less Memory
  - speeding up model-based OPs (with larger batch size)
  - enabling better cleaning/synthesis (with larger model size)

Optimization: Fault Tolerance

Opt 4: Batched Processing & Skipping

- Automatically skip problematic samples → **Worst-case guarantee**
  - Support batched processing for all OPs
  - Dynamic number of returned processed samples



Advanced Processing Hands-on

- Jupyter notebook part 4
  - how to efficiently process data with CUDA, OP fusion, ...

Tutorial Outline

- **Foundational Abilities**
  - Building Blocks of Data Processing: Data-Juicer's Operators
  - Composing Atomic Capabilities: Data-Juicer's Data Recipes

- **Advanced Data Processing**
  - Exploring Data Recipes: The Data-Juicer Sandbox Lab
  - From Exploration to Production: High-Performance Data Factory

- **Use Cases: From Text to Video Data Processing**

- **Resources and Conclusion**

Built-in Text Data Recipe

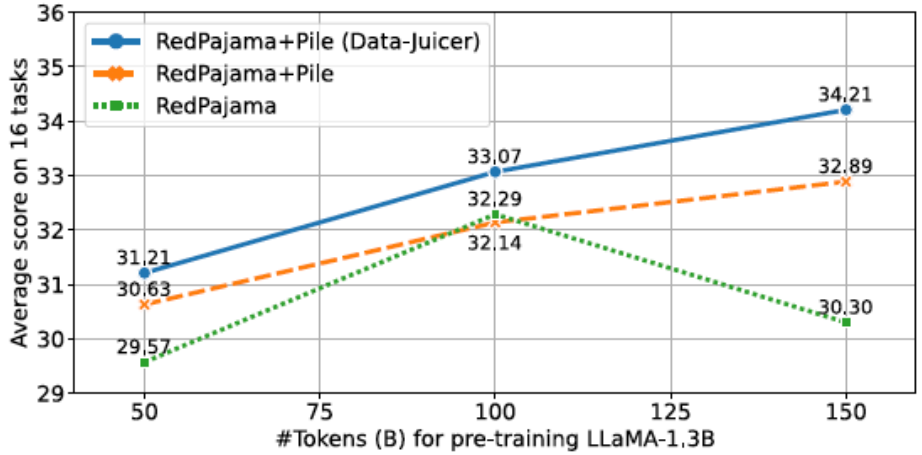
- How we build the data recipe?
  - Analyze all stats on the whole dataset
  - Set hyper-parameters of OPs with 3-sigma strategy

- Evaluation of the recipe: experiment settings

Exp Item	Pretraining	Finetuning (EN)	Finetuning (CN)
Base Model	LLaMA-1.3B	LLaMA-7B	LLaMA2-7B
Baseline Datasets	RedPajama & The Pile	Alpaca	Belle
Eval Benchmark	16 tasks in HELM	GPT-4 scoring	

Built-in Text Data Recipe

- Pretrain Performance of processed datasets
  - Consistently better performance
    - ✓ Refined datasets > Original datasets



- SFT Performance of processed datasets
  - English model:
    - ✓ Less than 4/5 samples → 17.5% higher winning rate
  - Chinese model:
    - ✓ Less than 1/10 samples → slightly higher winning rate

Model	Tuning Data	#Samples	Win	Tie
LLaMA-7B [33]	Alpaca	52k	16	100
	Data-Juicer	40k	**44**	
	Random (CFT, EN)	40k	19	105
	Data-Juicer	40k	**36**	
LLaMA2-7B (Chinese, FlagAlpha [41])	Belle	543k	28	99
	Data-Juicer	52k	**33**	
	Random (CFT, ZH)	52k	19	96
	Data-Juicer	52k	**45**	

Built-in Image & Video Data Recipes

- How we build the data recipes?
  - Quick experiments on single/multiple OPs via Data-Juicer Sandbox
  - Apply insights from sandbox:
    - ✓ Select top OPs and combine them, then repeatedly use good data

- Evaluation of the recipes: experiment settings

Exp Item	Image-to-text Task	Text-to-Video Task
Base Model	MGM-2B	T2V-Turbo
Baseline Datasets	MGM pretrain data (LLaVA + ALLaVA-4V)	Panda-70M, InternVid, MSR-VTT
Eval Benchmark	TextVQA, MMBench, MME	V-Bench

Built-in Image Data Recipe

- Performance of processed image dataset
  - Compared to original pre-train dataset: 1/10 distinct instances, 1/2 training instances
  - **Better performance** than baseline model, while **fewer data size (computation)**!

MGM-2B	Num. of Instances	Avg. Perf. Changes (%)	TextVQA	MMBench	MMBench-CN	MME-Perception	MME-Cognition
Official	1226k	-	56.2	59.8	-	1341	312
Reproduced	1226k	-	56.2	59.2	51.6	1334	302
Data-Juicer	**159k (x4)**	**+1.06**	54.41	62.08	52.32	1323.37	311.07

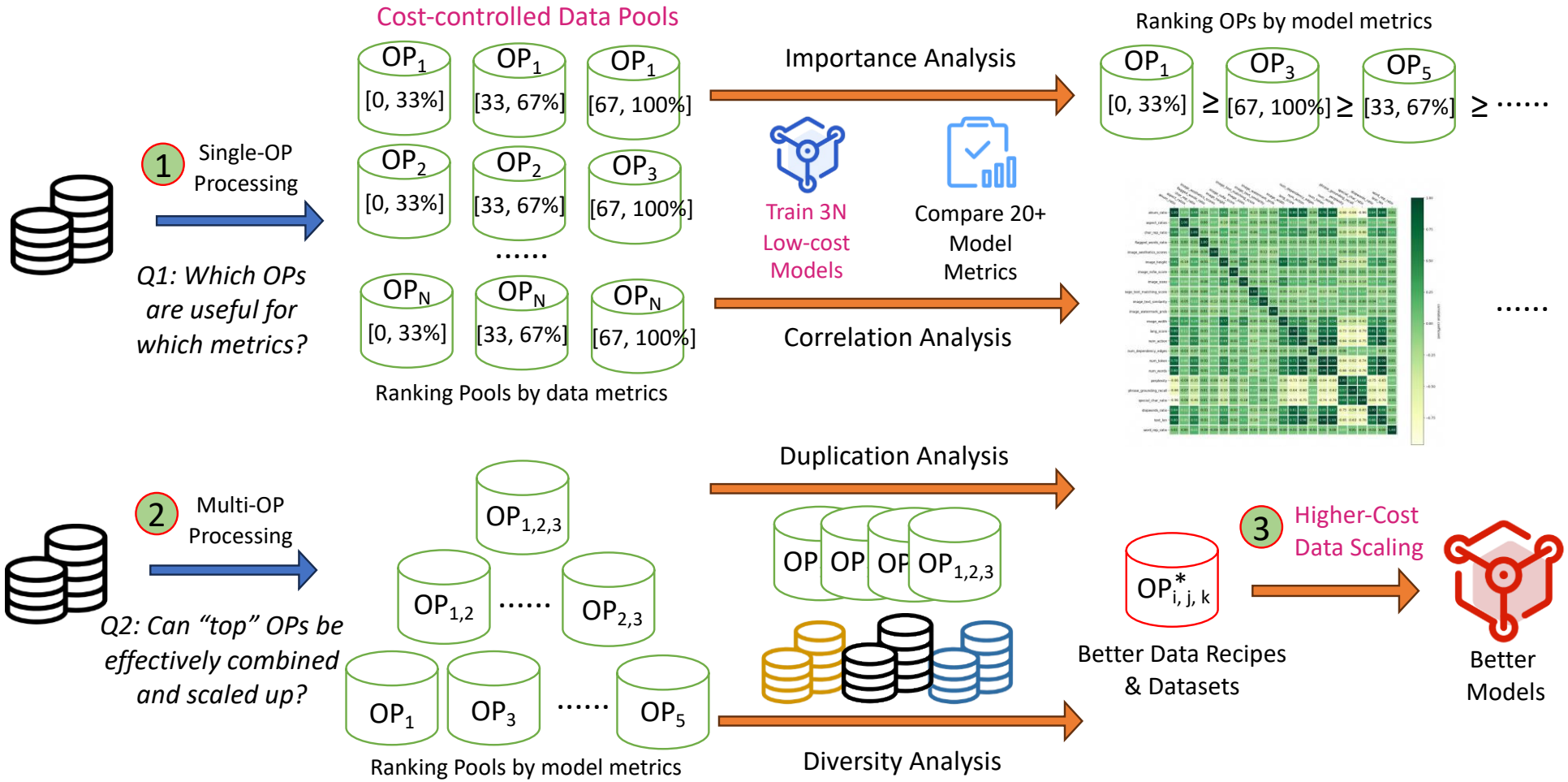
(repeat 4 times)

Built-in Video Data Recipe

- Performance of processed video datasets
  - Thanks to the probe-analyze-refine workflow in Data-Juicer Sandbox
  - Slight processing on data, **achieving a new SOTA performance!**

Models (Ranked by leaderboard)	Board Avg. (%)	Uniform Avg. (%)	Quality Avg. (%)	Semantic Avg. (%)
1. **Data-Juicer (T2V)**	**82.10**	**80.54**	83.14	**77.93**
2. VEnhancer (VC2) (He et al., 2024a)	81.97	80.00	**83.27**	76.73
3. LaVie-2 (Wang et al., 2023a)	81.75	79.50	83.24	75.76
4. T2V-Turbo (VC2) (Li et al., 2024)	81.01	78.67	82.57	74.76
5. Gen-2 (Esser et al., 2023)	80.58	80.11	82.47	77.75
6. VideoCrafter-2.0 (Chen et al., 2024b)	80.44	77.81	82.20	73.42
7. Pika Beta (2023-06) (Pika)	80.40	76.97	82.68	71.26
8. AnimateDiff-V2 (Guo et al., 2024)	80.27	76.33	82.90	69.75

Recap: Data-Juicer Sandbox



[3] (arXiv:2407.11784) Data-Juicer Sandbox: A Comprehensive Suite for Multimodal Data-Model Co-development

Sandbox Use Case: Studied OPs

OP Name	Modality	Statistics	Description
alphanumeric_filter	text	Alphanumeric Ratio	Alphanumeric ratio in the text.
character_repetition_filter	text	Character Repetition Ratio	Char-level n-gram repetition ratio in text.
flagged_words_filter	text	Flagged Word Ratio	Flagged-word ratio in the text
image_aesthetics_filter	image	Image Aesthetics Score	Aesthetics score of the image
image_aspect_ratio_filter	image	Image Aspect Ratio	Aspect ratio of the image
image_nsfw_filter	image	Image NSFW Score	NSFW score of the image
image_shape_filter	image	Image Width/Height	Width and height of the image
image_size_filter	image	Image Size	Size in bytes of the image
image_text_matching_filter	text, image	BLIP Image-Text Similarity	Image-text classification matching score based on a BLIP model
image_text_similarity_filter	text, image	CLIP Image-Text Similarity	Image-text feature cosine similarity based on a CLIP model
image_watermark_filter	image	Image Watermark Score	Predicted watermark score of the image based on an image classification model
language_id_score_filter	text	Language Score	Predicted confidence score of the specified language
perplexity_filter	text	Text Perplexity	Perplexity score of the text
phrase_grounding_recall_filter	text, image	Phrase Grounding Recall	Locating recall of phrases extracted from text in the image
special_characters_filter	text	Special Character Ratio	Special character ratio in the text
stopwords_filter	text	Stopword Ratio	Stopword ratio in the text
text_action_filter	text	Text Action Number	Number of actions in the text

Sandbox Use Case: Studied OPs

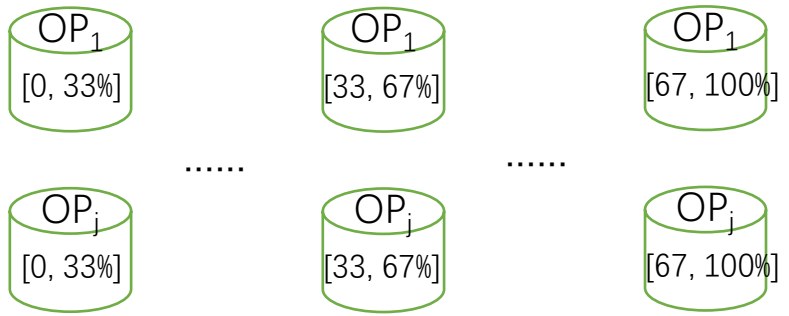
text_entity_dependency_filter	text	Entity Dependency Number	Number of dependency edges for an entity in the dependency tree of the text
text_length_filter	text	Text Length	Length of the text
token_num_filter	text	Token Number	Token number of the text
video_aesthetics_filter	video	Video Aesthetics Score	Aesthetics score of sampled frames in the video
video_aspect_ratio_filter	video	Video Aspect Ratio	Aspect ratio of the video
video_duration_filter	video	Video Duration	Duration of the video
video_frames_text_similarity_filter	text, video	Frames-Text Similarity	Similarities between sampled frames and text based on a CLIP/BLIP model
video_motion_score_filter	video	Video Motion Score	Motion score of the video
video_nsfw_filter	video	Video NSFW Score	NSFW score of the video
video_ocr_area_ratio_filter	video	Video OCR-Area Ratio	Detected text area ratio for sampled frames in the video
video_resolution_filter	video	Video Width/Height	Width and height of the video
video_watermark_filter	video	Video Watermark Score	Predicted watermark score of the sampled frames in the video based on an image classification model
words_num_filter	text	Word Number	Number of words in the text
word_repetition_filter	text	Word Repetition Ratio	Word-level n-gram repetition ratio in the text

Sandbox: Single-OP case

(compared to random & equal-sized data pool)

Task	OP-Generated Statistics	Average Performance Changes (%)		
		Data Pool (Low)	Data Pool (Mid)	Data Pool (High)
Image-to-Text	Image NSFW Score	7.13 ± 4.29	18.44 ± 18.45	**66.38 ± 32.65**
	Text Action Number	**59.90 ± 46.49**	0.29 ± 2.16	-2.05 ± 2.48
	Language Score	**49.90 ± 53.82**	0.85 ± 2.87	-1.43 ± 2.40
	CLIP Image-Text Similarity	1.20 ± 4.86	-1.81 ± 2.88	**49.81 ± 44.72**
	Phrase Grounding Recall	-0.49 ± 3.87	-0.58 ± 6.12	**49.39 ± 29.83**
	Image Width	**42.04 ± 57.27**	10.31 ± 12.59	1.35 ± 4.36
	Special Character Ratio	-3.08 ± 0.63	-0.75 ± 1.61	**39.67 ± 58.82**
	Flagged Word Ratio	**38.48 ± 27.76**	-0.39 ± 0.43	22.49 ± 29.81
	Image Height	**35.66 ± 48.62**	12.91 ± 10.42	18.73 ± 27.32
	Word Repetition Ratio	**33.14 ± 23.39**	2.59 ± 5.31	-0.55 ± 2.90
	Aesthetics Score	11.94 ± 12.21	**16.58 ± 25.70**	0.16 ± 3.67

(ranked by model performance changes)



Sandbox: Single-OP case

Task	OP-Generated Statistics	Average Performance Changes (%)		
		Data Pool (Low)	Data Pool (Mid)	Data Pool (High)
Text-related	Image NSFW Score	7.13 ± 4.29	18.44 ± 18.45	**66.38 ± 32.65**
	Text Action Number	**59.90 ± 46.49**	0.29 ± 2.16	-2.05 ± 2.48
	Language Score	**49.90 ± 53.82**	0.85 ± 2.87	-1.43 ± 2.40
Image-to-Text	CLIP Image-Text Similarity	1.20 ± 4.86	-1.81 ± 2.88	**49.81 ± 44.72**
	Phrase Grounding Recall	-0.49 ± 3.87	-0.58 ± 6.12	**49.39 ± 29.83**
	Image Width	**42.04 ± 57.27**	10.31 ± 12.59	1.35 ± 4.36
	Special Character Ratio	-3.08 ± 0.63	-0.75 ± 1.61	**39.67 ± 58.82**
	Flagged Word Ratio	**38.48 ± 27.76**	-0.39 ± 0.43	22.49 ± 29.81
	Image Height	**35.66 ± 48.62**	12.91 ± 10.42	18.73 ± 27.32
	Word Repetition Ratio	**33.14 ± 23.39**	2.59 ± 5.31	-0.55 ± 2.90
	Aesthetics Score	11.94 ± 12.21	**16.58 ± 25.70**	0.16 ± 3.67

Observation 1 (data v.s. model)

Generative models' efficacy is intimately tied to the fidelity of modalities they are trained to generate, which can be explicitly reflected in filtering processes on training data.

Sandbox: Single-OP case

Task	OP-Generated Statistics	Average Performance Changes (%)		
		Data Pool (Low)	Data Pool (Mid)	Data Pool (High)
Video-related	Video Aesthetics Score	-0.98 ± 0.08	0.13 ± 0.09	**0.96 ± 0.13**
	Video NSFW Score	**0.82 ± 0.36**	-0.05 ± 0.07	-0.57 ± 0.07
	Frames-Text Similarity	-1.45 ± 0.69	0.23 ± 0.21	**0.79 ± 0.15**
Text-to-Video	Special-Characters Ratio	**0.54 ± 0.36**	-0.13 ± 0.70	-0.14 ± 0.10
	Token Number	**0.53 ± 0.04**	0.18 ± 0.32	0.41 ± 0.25
	Video Height	-0.10 ± 0.21	0.12 ± 0.13	**0.46 ± 0.44**
	Video OCR-Area Ratio	**0.44 ± 0.04**	0.02 ± 0.63	-0.66 ± 0.23
	Word Number	-0.49 ± 0.07	-0.41 ± 0.72	**0.44 ± 0.45**
	Text Action Number	0.18 ± 0.56	-0.71 ± 0.28	**0.37 ± 0.28**
	Video Motion Score	-0.55 ± 0.40	**0.33 ± 0.21**	0.32 ± 0.15
	Video Aspect Ratio	-0.32 ± 0.14	**0.11 ± 0.18**	-0.02 ± 0.40
	Language Score	-0.21 ± 0.01	-0.03 ± 0.38	**0.09 ± 0.03**
	Video Duration	-0.58 ± 0.05	-0.16 ± 0.09	**0.04 ± 0.84**

Observation 1 (data v.s. model)

Generative models' efficacy is intimately tied to the fidelity of modalities they are trained to generate, which can be explicitly reflected in filtering processes on training data.

Sandbox: Single-OP case

Task	OP-Generated Statistics	Average Performance Changes (%)		
		Data Pool (Low)	Data Pool (Mid)	Data Pool (High)
Image-to-Text	Image NSFW Score	7.13 ± 4.29	18.44 ± 18.45	**66.38 ± 32.65**
	Text Action Number	**59.90 ± 46.49**	0.29 ± 2.16	-2.05 ± 2.48
	Language Score	**49.90 ± 53.82**	0.85 ± 2.87	-1.43 ± 2.40
	CLIP Image-Text Similarity	1.20 ± 4.86	-1.81 ± 2.88	**49.81 ± 44.72**
	Phrase Grounding Recall	-0.49 ± 3.87	-0.58 ± 6.12	**49.39 ± 29.83**
	Image Width	**42.04 ± 57.27**	10.31 ± 12.59	1.35 ± 4.36
	Special Character Ratio	-3.08 ± 0.63	-0.75 ± 1.61	**39.67 ± 58.82**
	Flagged Word Ratio	**38.48 ± 27.76**	-0.39 ± 0.43	22.49 ± 29.81
	Image Height	**35.66 ± 48.62**	12.91 ± 10.42	18.73 ± 27.32
	Word Repetition Ratio	**33.14 ± 23.39**	2.59 ± 5.31	-0.55 ± 2.90
	Aesthetics Score	11.94 ± 12.21	**16.58 ± 25.70**	0.16 ± 3.67

Observation 2 (diversity v.s. quality)

Image-to-text models place greater emphasis on data diversity, whereas text-to-video models prioritize data quality.

Sandbox: Single-OP case

Task	OP-Generated Statistics	Average Performance Changes (%)		
		Data Pool (Low)	Data Pool (Mid)	Data Pool (High)
Text-to-Video	Video Aesthetics Score	-0.98 ± 0.08	0.13 ± 0.09	**0.96 ± 0.13**
	Video NSFW Score	**0.82 ± 0.36**	-0.05 ± 0.07	-0.57 ± 0.07
	Frames-Text Similarity	-1.45 ± 0.69	0.23 ± 0.21	**0.79 ± 0.15**
	Special-Characters Ratio	**0.54 ± 0.36**	-0.13 ± 0.70	-0.14 ± 0.10
	Token Number	**0.53 ± 0.04**	0.18 ± 0.32	0.41 ± 0.25
	Video Height	-0.10 ± 0.21	0.12 ± 0.13	**0.46 ± 0.44**
	Video OCR-Area Ratio	**0.44 ± 0.04**	0.02 ± 0.63	-0.66 ± 0.23
	Word Number	-0.49 ± 0.07	-0.41 ± 0.72	**0.44 ± 0.45**
	Text Action Number	0.18 ± 0.56	-0.71 ± 0.28	**0.37 ± 0.28**
	Video Motion Score	-0.55 ± 0.40	**0.33 ± 0.21**	0.32 ± 0.15
	Video Aspect Ratio	-0.32 ± 0.14	**0.11 ± 0.18**	-0.02 ± 0.40
	Language Score	-0.21 ± 0.01	-0.03 ± 0.38	**0.09 ± 0.03**
	Video Duration	-0.58 ± 0.05	-0.16 ± 0.09	**0.04 ± 0.84**

Observation 2 (diversity v.s. quality)

Image-to-text models place greater emphasis on data diversity, whereas text-to-video models prioritize data quality.

Sandbox: Single-OP case

Task	OP-Generated Statistics	Average Performance Changes (%)		
		Data Pool (Low)	Data Pool (Mid)	Data Pool (High)
Image-to-Text	Image NSFW Score	7.13 ± 4.29	18.44 ± 18.45	**66.38 ± 32.65**
	Text Action Number	**59.90 ± 46.49**	0.29 ± 2.16	-2.05 ± 2.48
	Language Score	**49.90 ± 53.82**	0.85 ± 2.87	-1.43 ± 2.40
	CLIP Image-Text Similarity	1.20 ± 4.86	-1.81 ± 2.88	**49.81 ± 44.72**
	Phrase Grounding Recall	-0.49 ± 3.87	-0.58 ± 6.12	**49.39 ± 29.83**
	Image Width	**42.04 ± 57.27**	10.31 ± 12.59	1.35 ± 4.36
	Special Character Ratio	-3.08 ± 0.63	-0.75 ± 1.61	**39.67 ± 58.82**
	Flagged Word Ratio	**38.48 ± 27.76**	-0.39 ± 0.43	22.49 ± 29.81
	Image Height	**35.66 ± 48.62**	12.91 ± 10.42	18.73 ± 27.32
	Word Repetition Ratio	**33.14 ± 23.39**	2.59 ± 5.31	-0.55 ± 2.90
	Aesthetics Score	11.94 ± 12.21	**16.58 ± 25.70**	0.16 ± 3.67

Observation 3 (spatiotemporal dynamics)

Dynamic information in the data poses a heightened learning challenge for image-to-text generation compared to text-to-video generation.

Sandbox: Single-OP case

Task	OP-Generated Statistics	Average Performance Changes (%)		
		Data Pool (Low)	Data Pool (Mid)	Data Pool (High)
Text-to-Video	Video Aesthetics Score	-0.98 ± 0.08	0.13 ± 0.09	**0.96 ± 0.13**
	Video NSFW Score	**0.82 ± 0.36**	-0.05 ± 0.07	-0.57 ± 0.07
	Frames-Text Similarity	-1.45 ± 0.69	0.23 ± 0.21	**0.79 ± 0.15**
	Special-Characters Ratio	**0.54 ± 0.36**	-0.13 ± 0.70	-0.14 ± 0.10
	Token Number	**0.53 ± 0.04**	0.18 ± 0.32	0.41 ± 0.25
	Video Height	-0.10 ± 0.21	0.12 ± 0.13	**0.46 ± 0.44**
	Video OCR-Area Ratio	**0.44 ± 0.04**	0.02 ± 0.63	-0.66 ± 0.23
	Word Number	-0.49 ± 0.07	-0.41 ± 0.72	**0.44 ± 0.45**
	Text Action Number	0.18 ± 0.56	-0.71 ± 0.28	**0.37 ± 0.28**
	Video Motion Score	-0.55 ± 0.40	**0.33 ± 0.21**	0.32 ± 0.15
	Video Aspect Ratio	-0.32 ± 0.14	**0.11 ± 0.18**	-0.02 ± 0.40
	Language Score	-0.21 ± 0.01	-0.03 ± 0.38	**0.09 ± 0.03**
	Video Duration	-0.58 ± 0.05	-0.16 ± 0.09	**0.04 ± 0.84**

Observation 3 (spatiotemporal dynamics)

Dynamic information in the data poses a heightened learning challenge for image-to-text generation compared to text-to-video generation.

Sandbox: Single-OP case

Task	OP-Generated Statistics	Average Performance Changes (%)		
		Data Pool (Low)	Data Pool (Mid)	Data Pool (High)
Image-to-Text	Image NSFW Score	7.13 ± 4.29	18.44 ± 18.45	**66.38 ± 32.65**
	Text Action Number	**59.90 ± 46.49**	0.29 ± 2.16	-2.05 ± 2.48
	Language Score	**49.90 ± 53.82**	0.85 ± 2.87	-1.43 ± 2.40
	CLIP Image-Text Similarity	1.20 ± 4.86	-1.81 ± 2.88	**49.81 ± 44.72**
	Phrase Grounding Recall	-0.49 ± 3.87	-0.58 ± 6.12	**49.39 ± 29.83**
	Image Width	**42.04 ± 57.27**	10.31 ± 12.59	1.35 ± 4.36
	Special Character Ratio	-3.08 ± 0.63	-0.75 ± 1.61	**39.67 ± 58.82**
	Flagged Word Ratio	**38.48 ± 27.76**	-0.39 ± 0.43	22.49 ± 29.81
	Image Height	**35.66 ± 48.62**	12.91 ± 10.42	18.73 ± 27.32
	Word Repetition Ratio	**33.14 ± 23.39**	2.59 ± 5.31	-0.55 ± 2.90
	Aesthetics Score	11.94 ± 12.21	**16.58 ± 25.70**	0.16 ± 3.67

Observation 4 (modality alignment)

A high degree of match between different modalities within the data is crucial for model performance in both image-to-text and text-to-video generation.

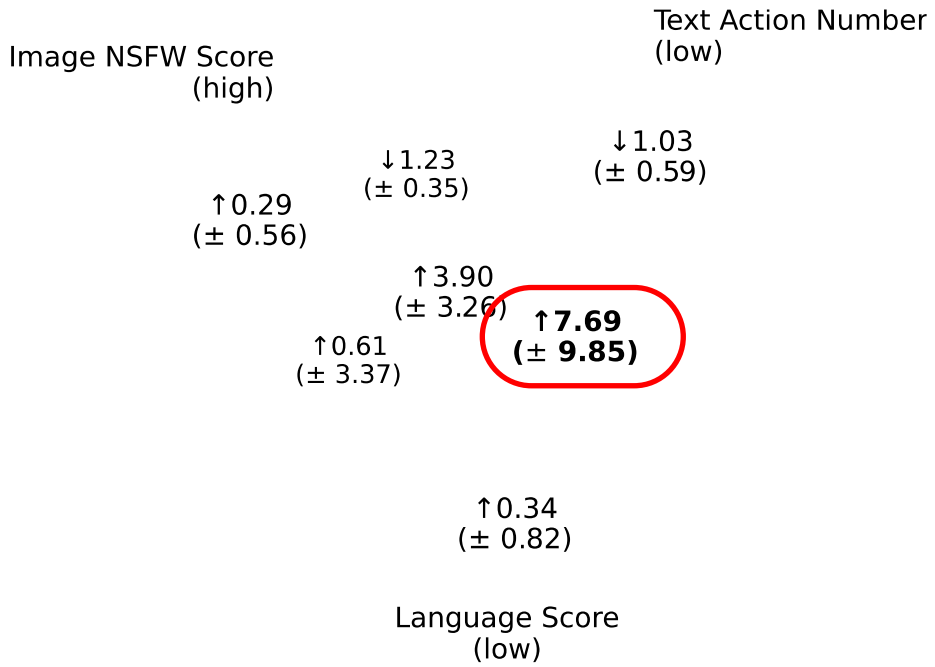
Sandbox: Single-OP case

Task	OP-Generated Statistics	Average Performance Changes (%)		
		Data Pool (Low)	Data Pool (Mid)	Data Pool (High)
Text-to-Video	Video Aesthetics Score	-0.98 ± 0.08	0.13 ± 0.09	**0.96 ± 0.13**
	Video NSFW Score	**0.82 ± 0.36**	-0.05 ± 0.07	-0.57 ± 0.07
	Frames-Text Similarity	-1.45 ± 0.69	0.23 ± 0.21	**0.79 ± 0.15**
	Special-Characters Ratio	**0.54 ± 0.36**	-0.13 ± 0.70	-0.14 ± 0.10
	Token Number	**0.53 ± 0.04**	0.18 ± 0.32	0.41 ± 0.25
	Video Height	-0.10 ± 0.21	0.12 ± 0.13	**0.46 ± 0.44**
	Video OCR-Area Ratio	**0.44 ± 0.04**	0.02 ± 0.63	-0.66 ± 0.23
	Word Number	-0.49 ± 0.07	-0.41 ± 0.72	**0.44 ± 0.45**
	Text Action Number	0.18 ± 0.56	-0.71 ± 0.28	**0.37 ± 0.28**
	Video Motion Score	-0.55 ± 0.40	**0.33 ± 0.21**	0.32 ± 0.15
	Video Aspect Ratio	-0.32 ± 0.14	**0.11 ± 0.18**	-0.02 ± 0.40
	Language Score	-0.21 ± 0.01	-0.03 ± 0.38	**0.09 ± 0.03**
	Video Duration	-0.58 ± 0.05	-0.16 ± 0.09	**0.04 ± 0.84**

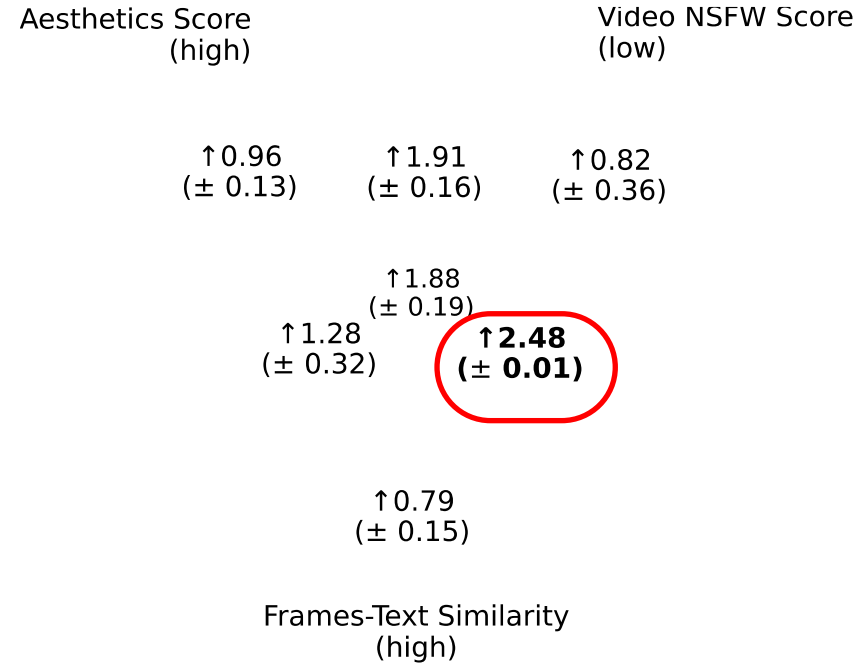
Observation 4 (modality alignment)

A high degree of match between different modalities within the data is crucial for model performance in both image-to-text and text-to-video generation.

Sandbox: Multi-OPs case



(a) Top-recipes for image-to-text model

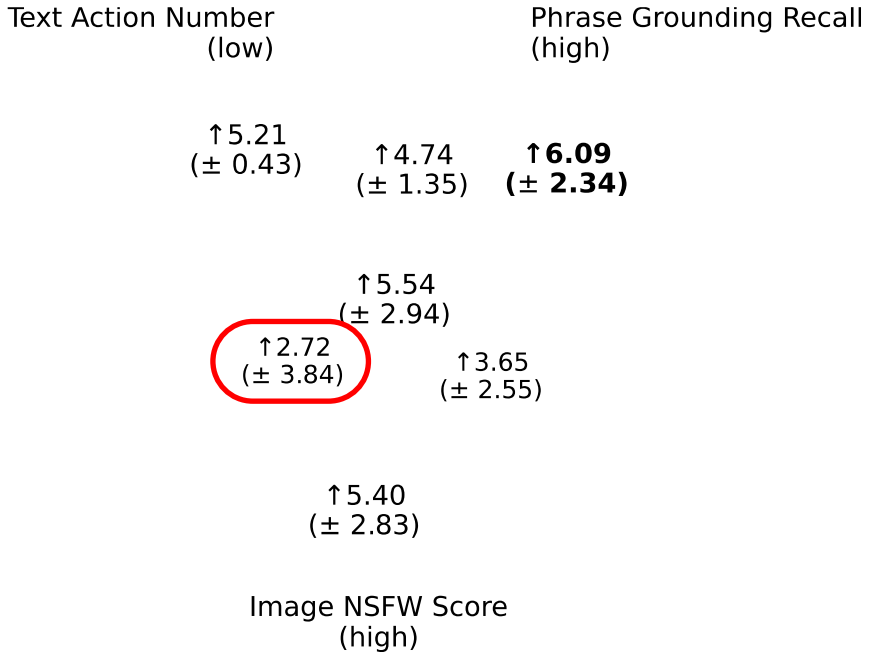


(b) Top-recipes for text-to-video model

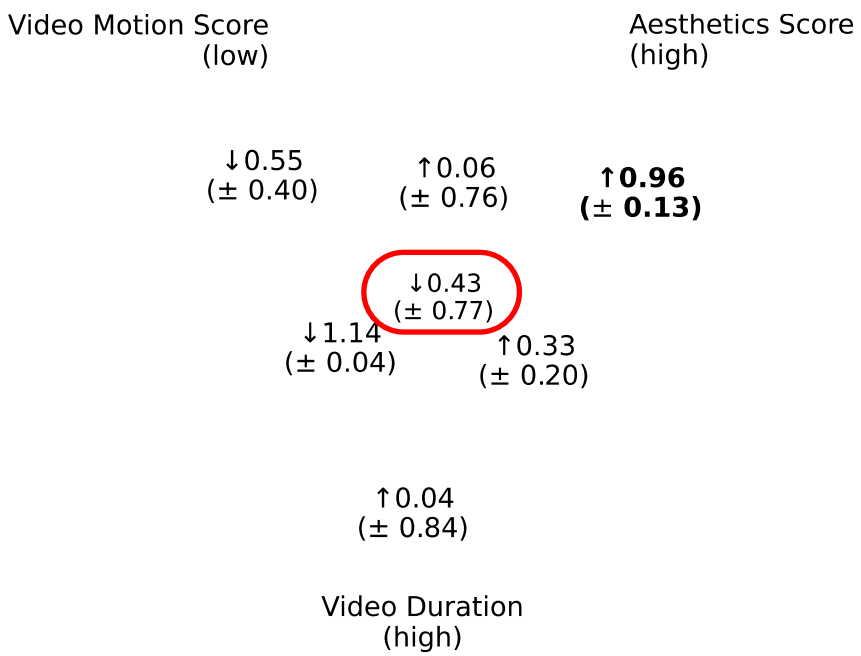
Observation 5 (1 + 1 < 2, sequentially)

The optimal data recipe does not necessarily result from combining the best individual OPs, nor does incorporating more high-performing OPs always lead to superior outcomes.

Sandbox: Multi-OPs case



(c) Cluster-recipes for image-to-text model

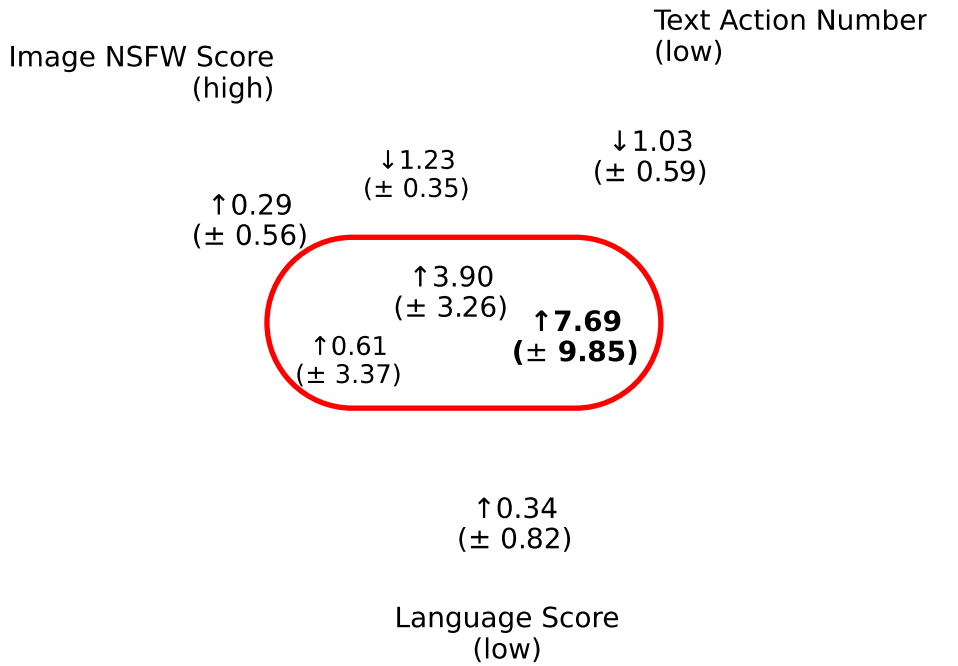


(d) Cluster-recipes for text-to-video model

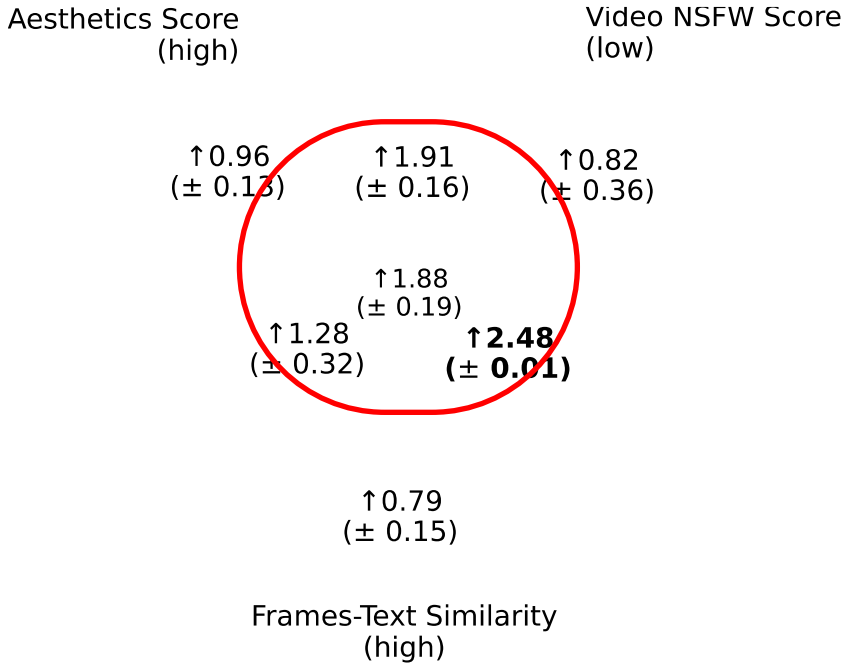
Observation 6 (1 + 1 < 2, orthogonally)

Combining OPs that excel in orthogonal dimensions on model or data does not guarantee complementary effects; rather, it is more likely that they will impede each other's performance.

Sandbox: Multi-OPs case



(a) Top-recipes for image-to-text model



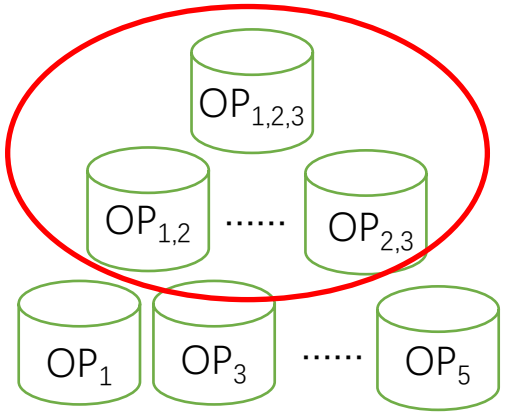
(b) Top-recipes for text-to-video model

Observation 7 (1 + x > x)

The performance of a single OP is positively correlated with the performance of the recipe created from its combination. Starting with high-performing OPs is a good initial step in exploring optimal higher-order data recipes.

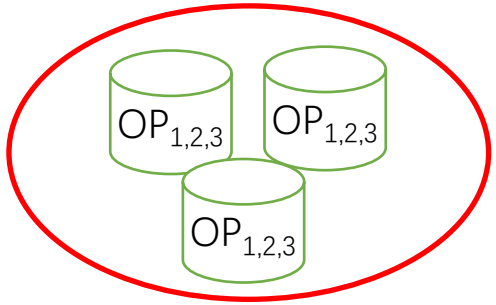
Sandbox: Data Duplication Case

A: Without Duplicates



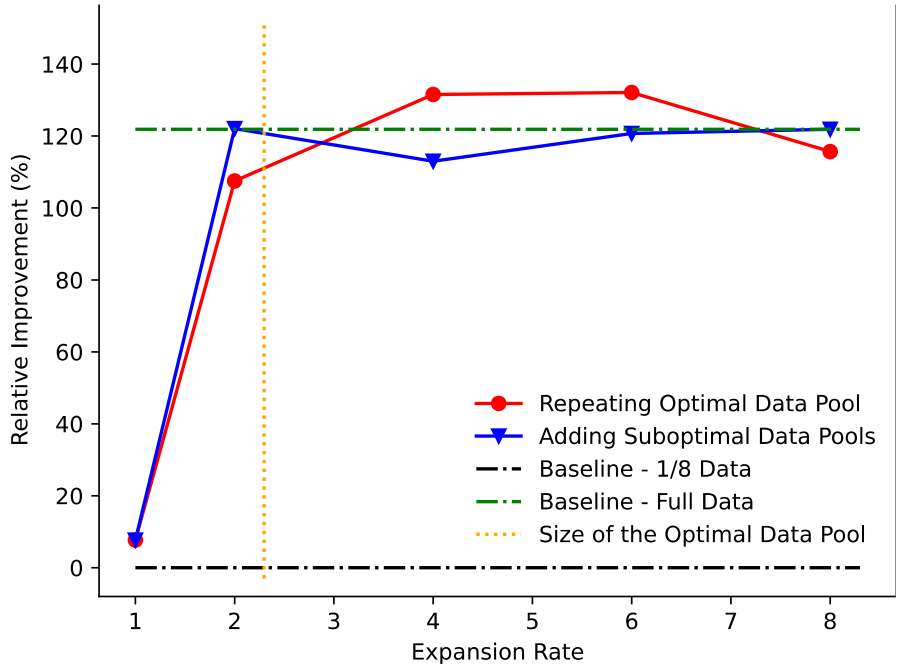
Ranking Pools by model metrics

B: With Duplication

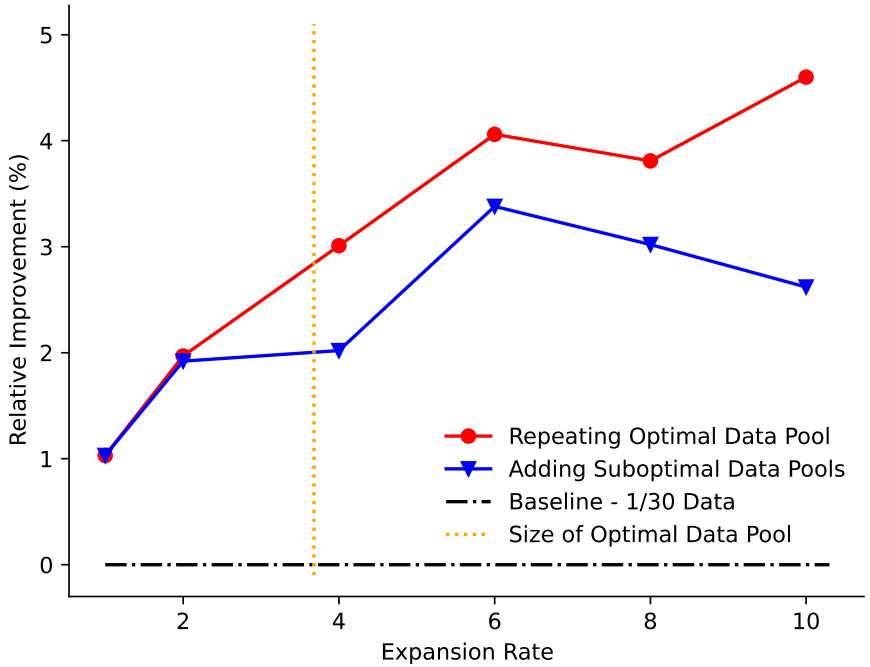


For text-only LLMs, A usually > B, but how about MLLMs?

Sandbox: Data Duplication Case



(a) Image-to-text



(b) Text-to-video

Observation 8 (effect of duplicates)

Duplicating high-quality data is beneficial for both image-to-text and text-to-video models. For the image-to-text model, optimal utilization of high-quality data may be achieved after four repetitions, while for the text-to-video model, it is effective to reuse high-quality data extensively between six to ten times.

The Use-cases Hands-on

- Jupyter notebooks part 5

  - 5.1: Synthetize data for Image-Text datasets

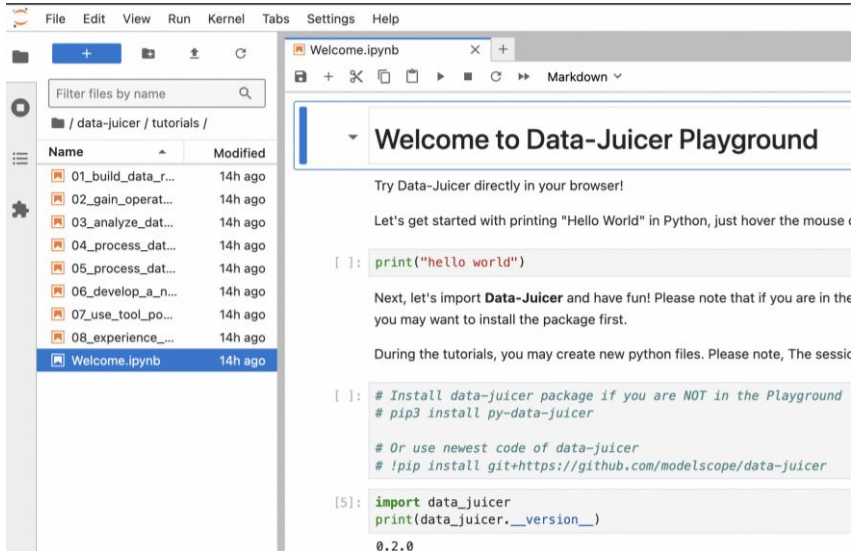
  - 5.2: Process video datasets from spacetime perspectives

Tutorial Outline

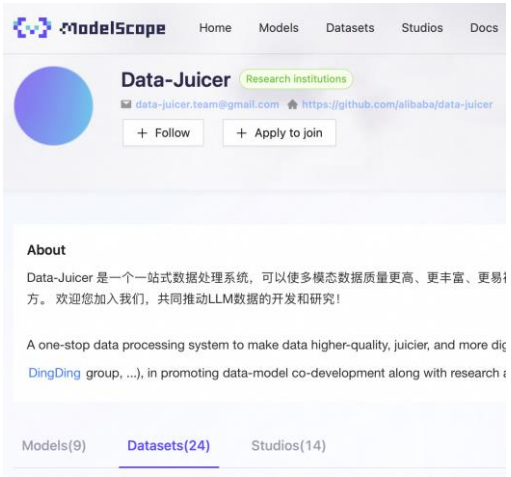
- **Foundational Abilities**
  - Building Blocks of Data Processing: Data-Juicer's Operators
  - Composing Atomic Capabilities: Data-Juicer's Data Recipes
- **Advanced Data Processing**
  - Exploring Data Recipes: The Data-Juicer Sandbox Lab
  - From Exploration to Production: High-Performance Data Factory
- Use Cases: From Text to Video Data Processing
- **Resources and Conclusion**

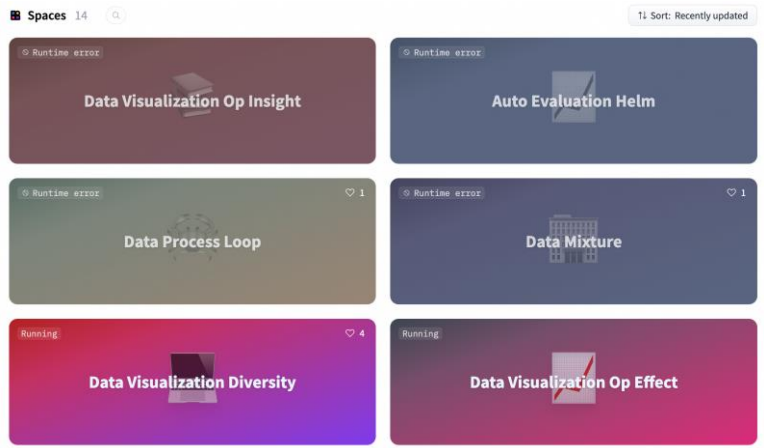
Interactive Resources

- Playgrounds
  - Jupyter Notebooks
  - Copilot Assistant



- 50+ interactive spaces, models, datasets on ModelScope & HuggingFace

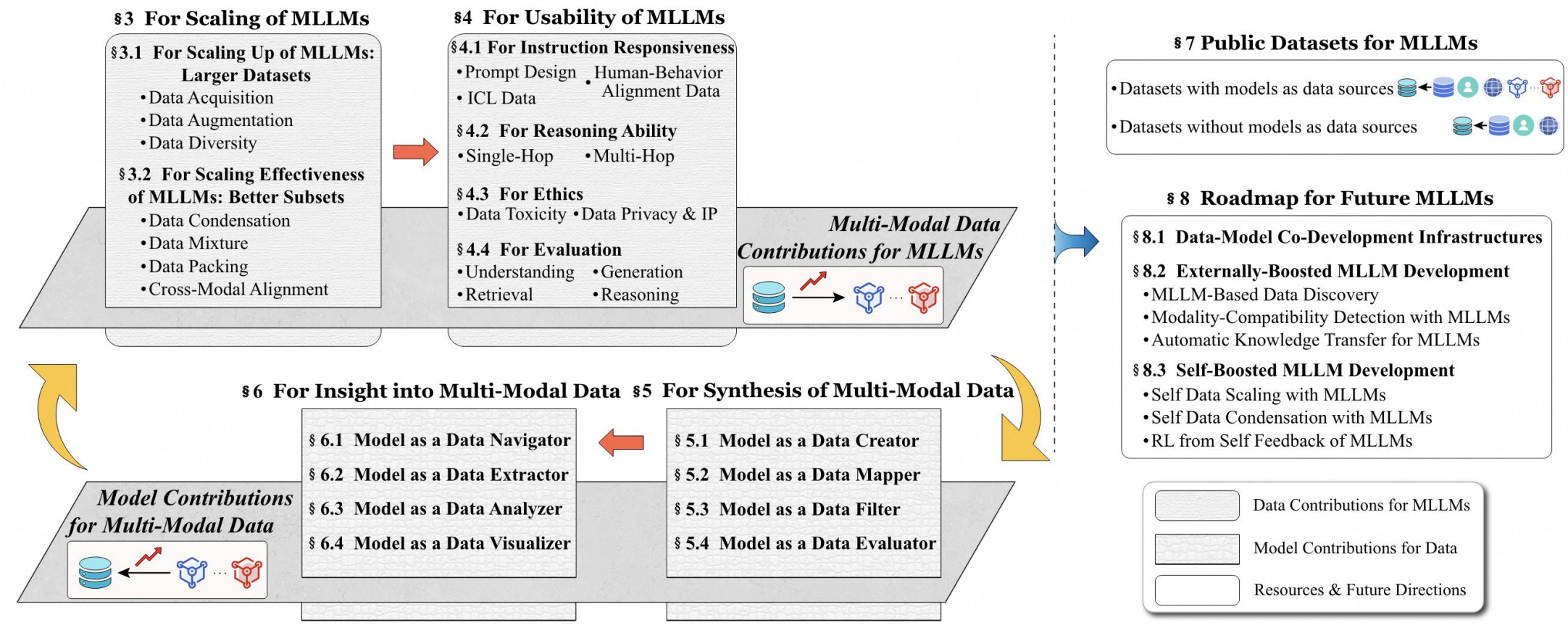




Data-Juicer Competitions

- 2000+ teams, open-sourced solutions
- Historical:
  - FT-Data Ranker: LLM finetuning data cleaning
  - Better Mixture: mix data from various sources
- On-going:
  - ModelScope-SORA: video generation
  - Better Synth: image understanding

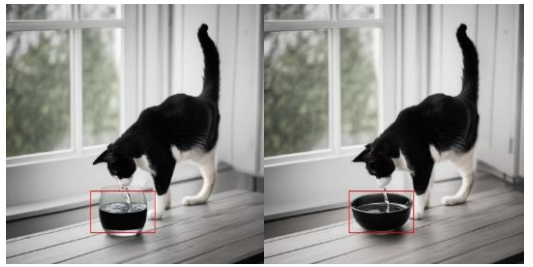
The Maintained Survey



[1] (arXiv:2407.08583) The Synergy between Data and Multi-Modal Large Language Models: A Survey from Co-Development Perspective
https://github.com/modelscope/data-juicer/docs/awesome_llm_data

Future Direction
MLLMs as Data Filter/Mapper

- Alleviate model collapse
  - MLLMs are generative models!
  - Inductive bias propagation → longer tailed distribution
- Better modality-compatibility
  - text-centric → any-to-any alignment
  - contrastive synthesis [7]
- Automatic compliance
  - privacy
  - licensing requirements



The left image shows a cat drinking water from **a glass**, while the right image shows the same cat drinking water from **a black bowl**. The difference is the type of container used for the water.

The difference between the two images lies in the object that the player is holding. In the left image, the player is holding **a baseball glove**, while in the right image, the player is holding **a basketball**.

[7] (arXiv:2408.04594) ImgDiff: Contrastive Data Synthesis for Vision Large Language Models

Future Direction
Data-Model Self-Boosting

- Dynamic Data & Model
  - Mutual enhancement

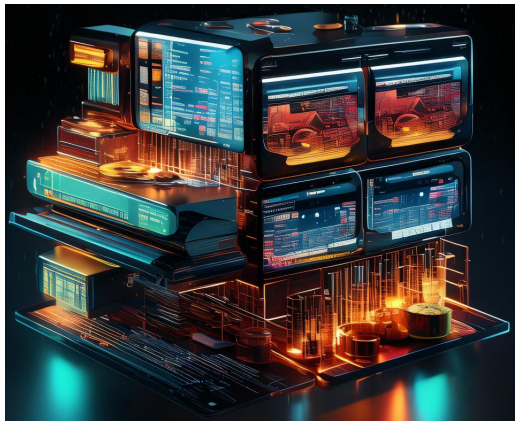
    e.g., [caption$_0$, fig$_0$] → train better MLLMs

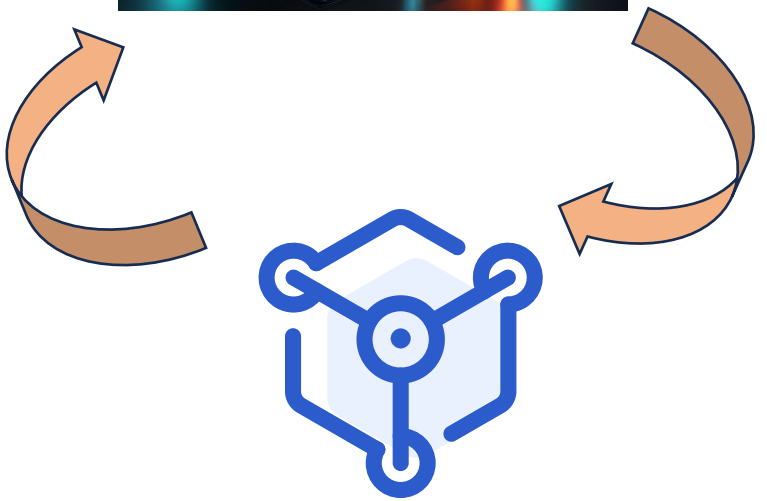
    → generate/clean better data:

    (more detailed, precise [caption$_1$, fig$_1$])

    → train better MLLMs → .....

- RL from self-feedback of data and MLLMs

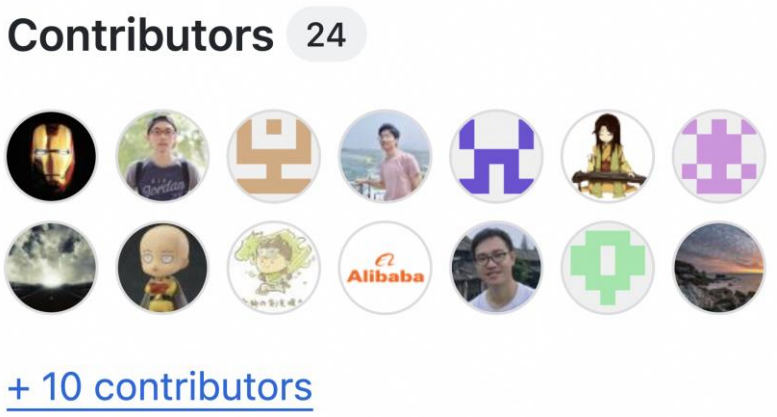




Conclusion

- Multi-modal Foundation Models
  - Scaling Laws (both computation and data)
  - "Garbage in, garbage out"

- Data-Juicer
  - open-sourced, effect-proven, industrial system
  - data analysis, cleaning, synthesis
  - sandbox: data-model co-development



- Under continuous maintenance, enhancement; welcome to join the community!

References

[1] (arXiv:2407.08583)

The Synergy between Data and Multi-Modal Large Language Models: A Survey from Co-Development Perspective

[2] (SIGMOD'24)

Data-juicer: A one-stop data processing system for large language models

[3] (arXiv:2405.14908)

Data Mixing Made Efficient: A Bivariate Scaling Law for Language Model Pretraining

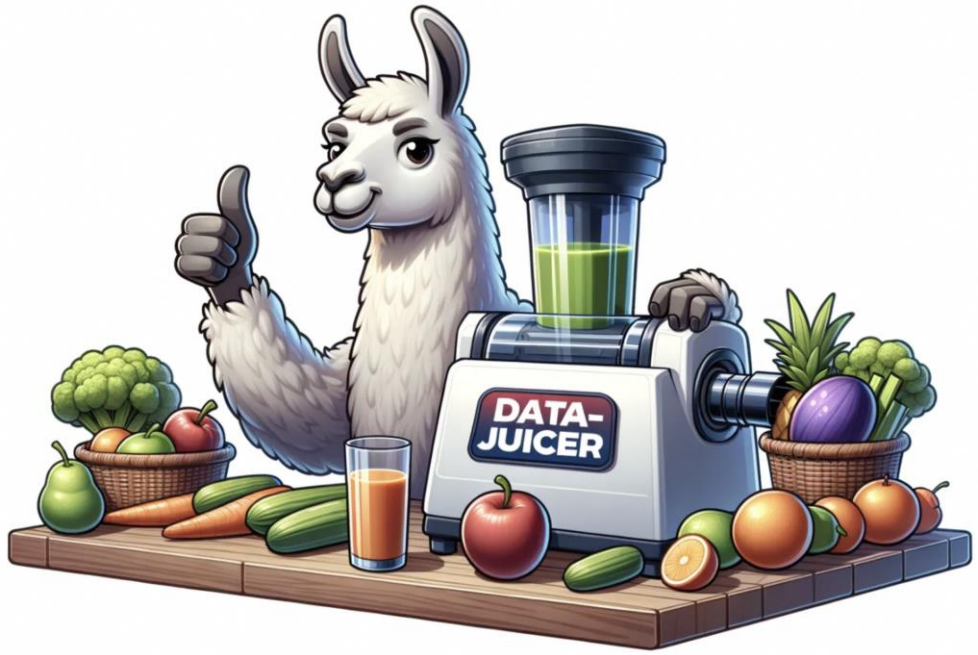
[4] (arXiv:2407.11784)

Data-Juicer Sandbox: A Comprehensive Suite for Multimodal Data-Model Co-development

[5] (arXiv:2408.04594)

ImgDiff: Contrastive Data Synthesis for Vision Large Language Models

Thanks!



GitHub: https://github.com/modelscope/data-juicer