# MaD-Mix: Multi-Modal Data Mixtures via Latent Space Coupling for Vision-Language Model Training

Wanyun Xie, Francesco Tonin, Volkan Cevher

Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)

DATA-FM @ ICLR 2026 Oral Presentation
April 26, 2026

# Outline

- Data processing pipeline for LLM/VLM training
- MaD-Mix: modality-aware alignment with latent-space coupling
- Empirical results of image-text tuning and tri-modal extension
- Conclusion and future direction

# Data processing pipeline for LLM/VLM training

- Modern LLMs (e.g., ChatGPT, Gemini, Claude, DeepSeek,...) employ the following general data pipeline:

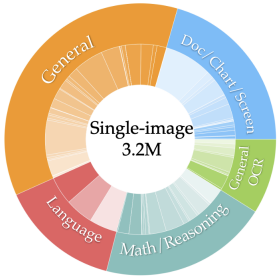
1. Data cleaning
   - Deduplication
   - Privacy and copyright
   - Quality filtering
   - Data age
2. Data selection
   - Data masking
   - Document packing
   - Data ordering
   - Synthetic data
3. Data mixing
   - Find optimal domain mixtures
   - Target evaluations
   - Online mixing approaches
   - Offline mixing approaches

# Data processing pipeline for LLM/VLM training

- Modern LLMs (e.g., ChatGPT, Gemini, Claude, DeepSeek,...) employ the following general data pipeline:

1. Data cleaning
   - Deduplication
   - Privacy and copyright
   - Quality filtering
   - Data age
2. Data selection
   - Data masking
   - Document packing
   - Data ordering
   - Synthetic data
3. Data mixing
   - Find optimal domain mixtures
   - Target evaluations
   - Online mixing approaches
   - Offline mixing approaches

# Data mixing in practice




| | | | |
|---|---|---|---|
| General (36.1%) | ALLaVA Inst (70.0K) [16] | AOKVQA (66.2 K) | Cambrian (filtered) (83.1 K) |
| CLEVR (0.7 K) | COCO Caption (20.0 K) | Hateful Memes (8.5 K) | IconQA (2.5 K) |
| Image Textualization (99.6 K) | LLaVA-158K (158.0 K) | LLaVA-Wild (train) (54.5 K) | LLaVAR (20.0 K) |
| OKVQA (9.0 K) | RefCOCO (50.6 K) | ScienceQA (5.0 K) | ShareGPT4o (57.3 K) |
| ShareGPT4V (91.0 K) | ST-VQA (17.2 K) | TallyQA (9.9 K) | Vision FLAN (186.1 K) |
| Visual7W (14.4 K) | VisText (10.0 K) | VizWiz (6.6 K) | VQARAD (0.3 K) |
| VQAv2 (82.8 K) | VSR (2.2 K) | WebSight (10.0 K) | InterGPS (1.3 K) |
| Doc/Chart/Screen (20.6%) | AI2D (GPT4V) (4.9 K) | AI2D (InternVL) (12.4 K) | AI2D (Original) (3.2 K) |
| Chart2Text (27.0 K) | ChartQA (18.3 K) | Diagram Image2Text (0.3 K) | Doc-VQA (10.2 K) |
| DVQA (20.0 K) | FigureQA (1.0 K) | HiTab (2.5 K) | Infographic VQA (4.4 K) |
| LRV Chart (1.8 K) | RoBUT SQA (8.5 K) | RoBUT WikiSQL (75.0 K) | RoBUT WTQ (38.2 K) |
| Screen2Words (15.7 K) | TQA (1.4 K) | UReader Caption (91.4 K) | UReader IE (17.3 K) |
| UReader KG (37.6 K) | UReader QA (252.9 K) | VisualMRC (3.0 K) | |
| Math/Reasoning (20.1%) | MAVIS MCollect (87.4 K) | MAVIS Data Engine (100.0 K) | Geo170K QA (67.8 K) |
| Geometry3K (2.1 K) | GEOS (0.5 K) | Geometry3K (MathV360K) (9.7 K) | GeoMVerse (MathV360K) (9.3 K) |
| GeoQA+ (MathV360K) (17.2 K) | MapQA (MathV360K) (5.2 K) | CLEVR-Math (5.3 K) | Geo170K Align (60.3 K) |
| MathQA (29.8 K) | Super-CLEVR (8.7 K) | TabMWP (45.2 K) | UniGeo (12.0 K) |
| GQA (72.1 K) | LRV Normal (10.5 K) | RAVEN (2.1 K) | Visual Genome (86.4K) |
| General OCR (8.9%) | ChromeWriting (8.8 K) | HME100K (74.5 K) | IIIT5K (2.0 K) |
| IAM (5.7 K) | K12 Printing (12.8 K) | OCR-VQA (80.0 K) | Rendered Text (10.0 K) |
| SynthDog-EN (40.1 K) | TextCaps (21.9 K) | TextOCR (25.1 K) | |
| Language (14.3%) | Magpie Pro (L3 MT) (150.0 K) | Magpie Pro (L3 ST) (150.0 K) | Magpie Pro (Qwen2 ST) (150.0 K) |

Figure: Data composition of LLaVA-One in the Visual Instruction Tuning stage [2].

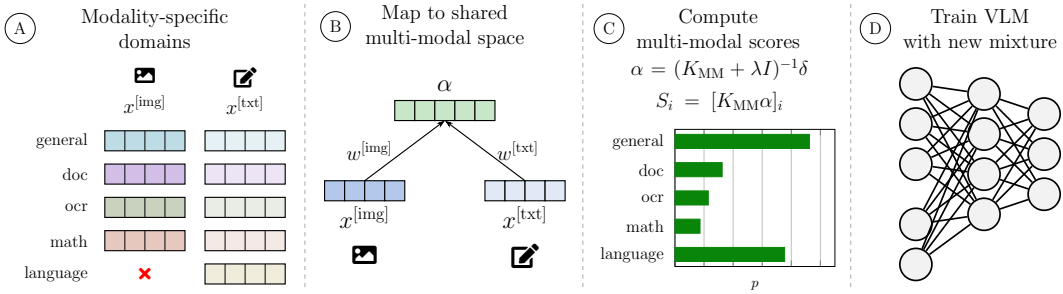
- Denote $k$ domains for pretraining as $D^{\text{train}} = \{D_1, D_2, \ldots, D_k\}$.
- The sampling distribution is $\boldsymbol{p} = (p_1, \ldots, p_k) \in \triangle^k$, the probability simplex over $k$ domains.
- *How to set $\boldsymbol{p}$ in an efficient way to improve performance?*

# Why is VLM data mixing hard?

- Many data mixing methods for LLMs, like DoReMi [4], DoGE [1], RegMix [3], Chameleon [5]...
- Data mixing in multi-modal settings still relies on manually tuning.
- Extending to multi-modality is challenging.
  - Multiple modalities: text, image, video
  - Missing modalities: pure text domains

| Domain | Text | Image | Video |
|---|---|---|---|
| General | ✓ | ✓ | ✗ |
| Doc / Chart | ✓ | ✓ | ✗ |
| Math / Reasoning | ✓ | ✓ | ✗ |
| Language | ✓ | ✗ | ✗ |
| VideoQA | ✓ | ✗ | ✓ |

# MaD-Mix pipeline




- Extract modality-specific domain embeddings from a mid-stage VLM.
- Couple domains in a shared latent space and compute multi-modal alignment scores.
- Softmax the scores to obtain training weights, then train the final VLM once.

## Single modality case

We equip each domain with a vector embedding $x_i \in \mathbb{R}^p$. Let $X = [x_1, \ldots, x_k]^\top \in \mathbb{R}^{k \times p}$.

### Single-modality alignment objective

For one modality, let $x_i \in \mathbb{R}^d$ be the embedding of domain $i$. We learn a shared direction:

$$\min_{w,e} \frac{1}{2\lambda} \sum_{i=1}^{k} e_i^2 + \frac{1}{2}\|w\|_2^2 \qquad \text{s.t. } e_i = 1 - w^\top x_i, \;\; i = 1, \ldots, k.$$

The score of domain $i$ is then

$$S_i = x_i^\top w_\star.$$

**Remarks:**

- $x_i$ lies along the common direction shared by many domains $\Rightarrow S_i$ is **large**.
- $x_i$ is hard to reconstruct from other domains $\Rightarrow S_i$ is **small**.
- Interpretation: $S_i$ measures how well domain $i$ represents the *shared geometry* of the dataset.

## Alignment score computation

- Primal problem of alignment score of domain $i$:

$$\boxed{\text{P}} \quad \min_{w,e} \frac{1}{2}\sum_{i=1}^{k} e_i^2 + \frac{\lambda}{2}\|w\|^2 \quad \text{s.t. } e_j = 1 - w^\top x_i \ \forall i$$

- Lagrangian:

$$\mathcal{L}(\boldsymbol{w}, \boldsymbol{e}; \boldsymbol{\alpha}) = \frac{1}{2\lambda}\sum_{i=1}^{k} e_i^2 + \frac{1}{2}\|w\|^2 - \sum_{i=1}^{k} \alpha_i (e_i - 1 + w^\top x_i).$$

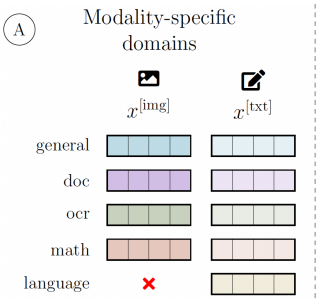
- Elimination of $\boldsymbol{w}, \boldsymbol{e}$ in the optimality conditions gives:

$$\boxed{\text{D}} \quad \boldsymbol{\alpha} = (K + \lambda \boldsymbol{I})^{-1} 1_k,$$

  with linear kernel matrix $K = [x_i^\top x_j]_{i,j=1}^k$.

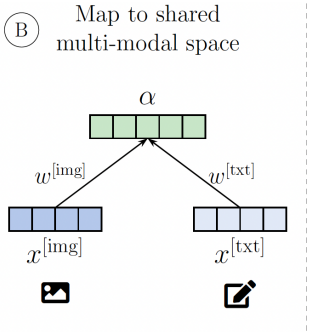
- The score of domain $i$: $S_i = [K(K + \lambda I)^{-1} 1_k]_i$.

# From single-modality to multi-modality




- Extract embeddings for each modality independently:
  $x_i^{[v]}$ denote the embedding of the modality $v$ in domain $i$.

# From single-modality to multi-modality



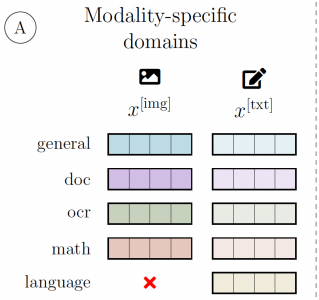

- Extract embeddings for each modality independently:
  $x_i^{[v]}$ denote the embedding of the modality $v$ in domain $i$.

- Map different modalities to a shared space $\alpha$ $\rightarrow$ **cross-modal coupling**.

$$J = \frac{1}{2\lambda}\sum_{i=1}^{k}(1-(w^{[v]})^\top x_i^{[v]})^2 + \frac{1}{2}\|w^{[v]}\|^2$$

$$\geq \sum_{i=1}^{k}(1-(w^{[v]})^\top x_i^{[v]})\,\alpha_i - \frac{\lambda}{2}\|\alpha\|^2 + \frac{1}{2}\|w^{[v]}\|^2$$

# Handle missing modalities




- Some domains don't include all modalities.
- **Multi-modal objective:**

$$J_{\text{MM}} = \sum_{v=1}^{V} \sum_{i=1}^{k} \left[ \left( \delta_i^{[v]} - (w^{[v]})^\top x_i^{[v]} \right) \alpha_i - \frac{\lambda}{2} \alpha_i^2 \right] + \frac{1}{2} \sum_{v=1}^{V} \|w^{[v]}\|^2$$

where $\delta_i^{[v]} \in \{0, 1\}$ is the modality indicator.

- Set $x_i^{[v]} = 0$ and $\delta_i^{[v]} = 0$ if a modality is missing.
- Missing modalities **do not contribute** to $J_{\text{MM}}$.

## From latent-space coupling to data weights

- Shared latent variables

$$\alpha = (K_{\mathrm{MM}} + \lambda I)^{-1}\delta$$

- Modality-aware domain score

$$S_i^{[v]} = \left[K^{[v]}(K_{\mathrm{MM}} + \lambda I)^{-1}\delta\right]_i$$

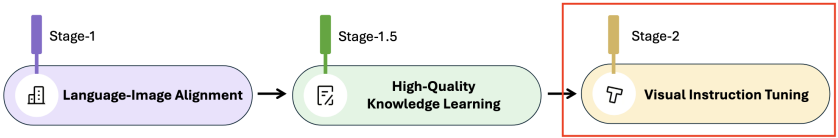
- Training mixture

$$p_i = \frac{\exp\left(\sum_{v=1}^{V} S_i^{[v]}\right)}{\sum_j \exp\left(\sum_{v=1}^{V} S_j^{[v]}\right)}$$

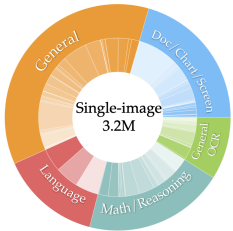
- $K_{\mathrm{MM}} = \sum_v K^{[v]}$ couples all modalities.
- $\delta_i$ counts how many modalities exist in domain $i$.
- Missing modalities contribute zero rather than noise.

## Experimental setup and baselines

- Our experiments focus on the Visual Instruction Tuning stage in the LLaVA-One [2].



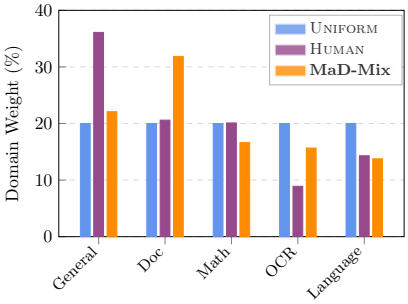

- LLaVA-One data includes 5 domains.
  - Language: pure text data
  - Others: image-text pairs

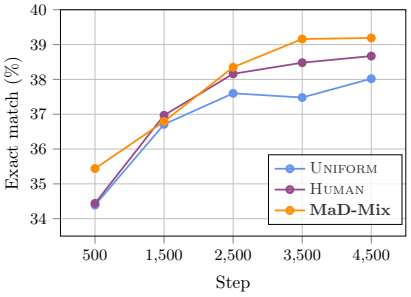



- Baselines
  - UNIFORM: Equal weights
  - HUMAN: Hand-tuned mixture (reported in [2])
  - AVG: Average of unimodal weights
  - FUSED: Weights from early-fused embeddings

# Image-text instruction tuning

- **MaD-Mix** beats UNIFORM with 56% of the steps.
- **MaD-Mix** matches HUMAN with 78% of the steps.



## Image-text instruction tuning

- Better than AVG: treating modalities independently loses cross-modal structure.
- Better than FUSED: early fusion can blur domain-specific modality effects.
- Domain weights can transfer from LLaVA-0.5B to LLaVA-7B.

| Benchmark | UNIFORM | HUMAN | AVG | FUSED | **MaD-Mix** |
|---|---|---|---|---|---|
| AI2D | 42.78 | 43.75 | 45.50 | 44.59 | 43.52 |
| DocVQA | 42.90 | 42.66 | 42.44 | 42.67 | 42.92 |
| InfoVQA | 22.25 | 22.61 | 22.43 | 23.50 | 22.13 |
| MathVerse | 18.27 | 17.26 | 18.32 | 19.29 | 18.91 |
| MMBench | 36.34 | 40.21 | 39.86 | 37.71 | 42.44 |
| MMStar | 33.45 | 36.04 | 33.50 | 34.44 | 35.88 |
| MMMU | 30.00 | 29.67 | 29.00 | 29.22 | 29.78 |
| ScienceQA | 62.42 | 65.84 | 64.80 | 63.46 | 64.50 |
| OCRBench | 45.30 | 44.60 | 45.30 | 43.50 | 45.80 |
| RealworldQA | 46.27 | 44.05 | 45.49 | 45.36 | 46.54 |
| Average | 38.00 | 38.67 | 38.66 | 38.37 | **39.24** |
| Number over UNIFORM | - | 5/10 | 6/10 | 6/10 | 8/10 |

Table: Data mixing strategies for **LLaVA-0.5B** image-text instruction tuning.

## Image-text instruction tuning

- Better than AVG: treating modalities independently loses cross-modal structure.
- Better than FUSED: early fusion can blur domain-specific modality effects.
- Domain weights can transfer from LLaVA-0.5B to LLaVA-7B.

| Benchmark | UNIFORM | HUMAN | AVG | FUSED | **MaD-Mix** |
|---|---|---|---|---|---|
| AI2D | 74.48 | 74.03 | 75.10 | 75.74 | 75.58 |
| DocVQA | 57.91 | 58.64 | 58.28 | 57.29 | 58.32 |
| InfoVQA | 34.76 | 35.91 | 36.95 | 36.06 | 36.23 |
| MathVerse | 29.31 | 26.85 | 27.33 | 28.68 | 28.55 |
| MMBench | 75.69 | 76.12 | 76.23 | 75.77 | 75.74 |
| MMStar | 49.04 | 50.26 | 50.44 | 49.46 | 50.19 |
| MMMU | 46.33 | 46.78 | 46.78 | 46.78 | 46.89 |
| ScienceQA | 87.31 | 90.38 | 89.53 | 85.52 | 90.23 |
| OCRBench | 56.80 | 57.30 | 56.70 | 56.60 | 57.90 |
| RealworldQA | 58.17 | 57.91 | 56.99 | 57.65 | 57.47 |
| Average | 56.98 | 57.42 | 57.43 | 56.96 | **57.71** |
| Number over UNIFORM | - | 7/10 | 7/10 | 5/10 | 8/10 |

Table: Transfer weights from LLaVA-0.5B to **LLaVA-7B** for image-text instruction tuning.

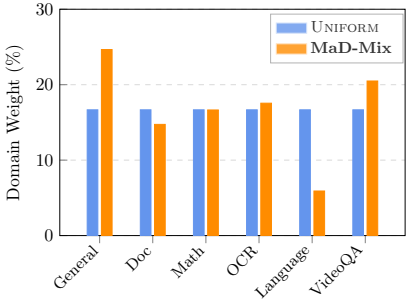
## Computational cost is negligible

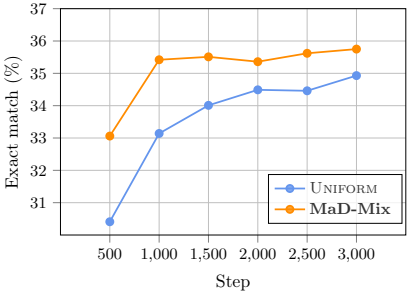
- Embedding extraction is an inference pass of the pretrained checkpoint from mid-stage training.
- The computational complexity of score $S$ is $\mathcal{O}(k^3)$, and remains fast in practice since $k$ is small.
- Computation overhead is $<1$ GPU-hour, which is marginal compared with manually tuning.

| **Component** | **Cost (h)** |
|---|---|
| Embedding extraction | $0.58$ |
| Score computation | $0.01$ |
| Total | $0.59$ |
| Training (0.5B) | $90$ |
| Training (7B) | $620$ |

## Extend to video-image-text training

- Add VideoQA domain (video-text pair data) in, and add two video benchmarks.




| | 0.5B | | | | 7B | | | |
|---|---|---|---|---|---|---|---|---|
| | Uniform | Avg | Fused | **MaD-Mix** | Uniform | Avg | Fused | **MaD-Mix** |
| Average | 34.93 | 34.53 | 34.74 | **35.75** | 52.91 | 53.39 | 52.88 | **54.40** |
| # over Unif. | - | 4/12 | 7/12 | 9/12 | - | 6/12 | 5/12 | 10/12 |

# Conclusion

## Summary of MaD-Mix

- **MaD-Mix** turns VLM mixture design into a lightweight latent-space alignment problem.
- **Handle missing modalities**, enabling joint optimization of language-only and multimodal domains.
- Improve both **accuracy** and **training efficiency** in both image-text and tri-modal settings.
- Reduce data curation cost from expensive manual tuning to **negligible** levels.

## Future direction

- Online method - How to adjust domain weights during VLM training?
- Domain definition - Will the domain construction affect domain weights?
- Modality interactions - Better cross-modal modeling may further improve data mixing.

## Thank you

Questions?

MaD-Mix: Multi-Modal Data Mixtures via Latent Space Coupling for Vision-Language Model Training

wanyun.xie@epfl.ch, francesco.tonin@epfl.ch

# References I


[1] Simin Fan, Matteo Pagliardini, and Martin Jaggi. DOGE: Domain reweighting with generalization estimation. In *International Conference on Machine Learning*, 2024. (Cited on page 6.)

[2] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. LLaVA-Onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. (Cited on pages 5 and 14.)

[3] Qian Liu, Xiaosen Zheng, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, and Min Lin. Regmix: Data mixture as regression for language model pre-training. *arXiv preprint arXiv:2407.01492*, 2024. (Cited on page 6.)

[4] Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V. Le, Tengyu Ma, and Adams Wei Yu. DoReMi: Optimizing data mixtures speeds up language model pretraining. In *Advances in Neural Information Processing Systems*, 2023. (Cited on page 6.)

[5] Wanyun Xie, Francesco Tonin, and Volkan Cevher. Chameleon: A flexible data-mixing framework for language model pretraining and finetuning. In *Forty-second International Conference on Machine Learning*, 2025. (Cited on page 6.)

# Comparison of Orthogonal and Alignment scores

Table: **Comparison of Orthogonal and Alignment scores for LLaVA-0.5B image-text instruction tuning.** Orthogonal score is even worse than UNIFORM (37.62 vs 38.00).

| **Benchmark** | Orthogonal score | Alignment score (**MaD-Mix**) |
|---|---|---|
| AI2D | 43.04 | 43.52 |
| DocVQA | 41.58 | 42.92 |
| InfoVQA | 21.49 | 22.13 |
| MathVerse | 15.99 | 18.91 |
| MMBench | 32.65 | 42.44 |
| MMStar | 34.52 | 35.88 |
| MMMU | 30.22 | 29.78 |
| ScienceQA | 64.25 | 64.50 |
| OCRBench | 46.30 | 45.80 |
| RealworldQA | 46.14 | 46.54 |
| Average | 37.62 | **39.24** |

# Domain weights transfer from LLaVA-0.5B to Qwen2-VL

Table: **Transfer domain weights from LLaVA-0.5B to Qwen2-VL-2B for video-image-text instruction tuning.**

| Benchmark | UNIFORM | AVG | FUSED | **MaD-Mix** |
|---|---|---|---|---|
| AI2D | 67.78 | 67.94 | 67.29 | 68.26 |
| DocVQA | 75.10 | 75.76 | 80.11 | 78.11 |
| InfoVQA | 42.69 | 42.42 | 42.72 | 44.02 |
| MathVerse | 21.19 | 18.78 | 23.73 | 23.98 |
| MMBench | 59.02 | 56.44 | 55.33 | 59.71 |
| MMStar | 41.37 | 42.90 | 40.89 | 41.11 |
| MMMU | 37.44 | 36.00 | 35.98 | 37.44 |
| ScienceQA | 77.89 | 79.13 | 77.84 | 79.23 |
| OCRBench | 71.60 | 73.80 | 72.90 | 72.30 |
| RealworldQA | 58.82 | 58.82 | 58.04 | 58.69 |
| Video_MMMU | 21.02 | 21.50 | 20.32 | 20.83 |
| MVBench | 56.38 | 56.88 | 56.88 | 56.92 |
| Average | 52.53 | 52.53 | 52.67 | **53.38** |
| Number over UNIFORM | - | 7/12 | 6/12 | 8/12 |