MergeMix: A Unified Augmentation Paradigm for Visual and Multi-Modal Understanding

Xin Jin[1,*], Siyuan Li[1,2,*], Siyong Jian[1], Kai Yu[1], Huan Wang[1,✉]

[1] Westlake University, [2] Zhejiang University

The 14th International Conference on Learning Representations

Introduction of MergeMix

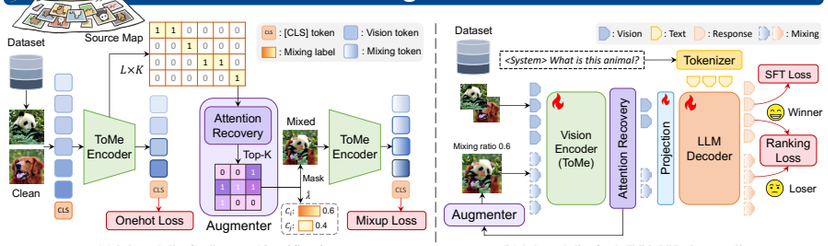
1. Motivation

- Hard to achieve an optimal **trade-off** between the efficiency and performance of mixup.
- How to **extending mixup** to MLLMs from classical image corruptions to data-dependent samples.

2. Core Ideas & Results

- Token Merge (ToMe) for **accelerating** training & evaluation, BSM strategy for preserving contextual features.
- Preference tuning for **better aligning** the visual to the language space with raw & augmented images.
- MergeMix achieves **SOTA** on several image classification datasets and benchmarks, demonstrating the advantages of the training paradigm across various MLLM benchmarks.

Method I: Image Classification



(a) MergeMix for Image Classification (b) MergeMix for MLLM Understanding

Fig 1. The overall of the two scenarios of MergeMix. *Left*: Image Classification & *Right*: MLLMs.

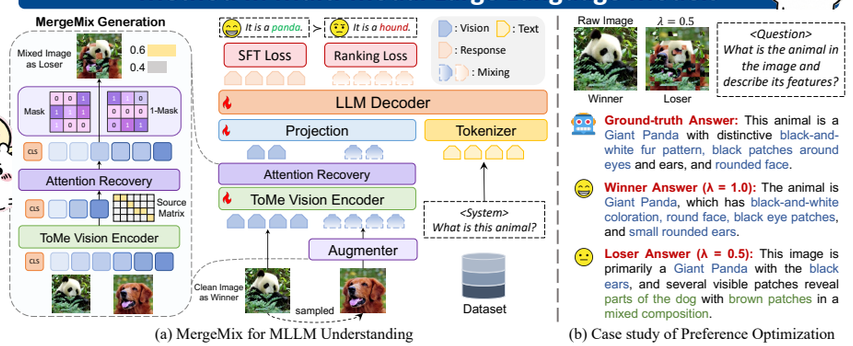
Step 1: Image Mixing via Token Merge

Sample Mixing. A ViT-based encoder (w. ToMe) obtain the output Z_K, attention map $A_K \in K\times K$ and source matrix $S \in L\times K$:

$$S, A_K, Z_K = \text{ToMeAttention}(Z_L, r).$$

Generating Mixing Mask. Recover the attention score by the matrix by $\hat{A}_L = R_{K\to L}(A_K, S)$, generate the mask $\mathcal{M}$ by TopK$(\cdot)$ from the recovered attention map.

Method II: Multi-modal Large Language Models



(a) MergeMix for MLLM Understanding (b) Case study of Preference Optimization

Fig 2. (a) Overall illustration of MergeMix for MLLM, (b) Case study of on LLaVA-v1.5-7B.

Step 2: A Unified Augmentation Paradigm, CLS. to MLLMs

Label Mixing. Re-scaling the mixing ratio $\hat{\lambda}$ by the ratio λ in the mask $\mathcal{M}$ (as μ) and the merge ratio r (as σ):

$$\hat{\lambda} \sim \mathcal{N}(\mu, \sigma), \quad \lambda = \text{clip}\left(\frac{\lambda - \min(\lambda)}{\max(\lambda) - \min(\lambda) + \tau}, 0, 1\right).$$

Reformulate Loss Function. $\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{SFT}} + \mathcal{L}^{\text{Mix}}_{\text{SimPO}}$ / $\mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{MCE}}$.

$$\mathcal{L}^{\text{Mix}}_{\text{SimPO}} = -\mathbb{E}_{(x,\hat{x},y)\sim D}[\log\sigma(\frac{\beta}{|y|}\log\pi_\theta(y \mid x) - \frac{\beta}{|y|}\log\pi_\theta\{y \mid \hat{x}) - (1 - \hat{\lambda}))].$$

Main Experiments

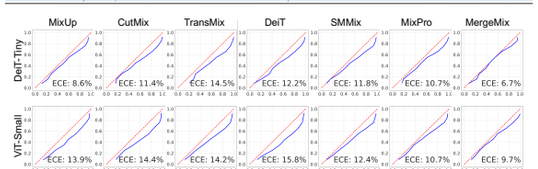
Tab 1 & 2. Top-1 accuracy (%) ↑ of mixup methods on CIFAR100 and Stanford-Cars datasets.

Method	DeiT-T	DeiT-S	ViT-S	ViT-B	ViT-L
Vanilla	64.70	65.81	62.64	63.33	61.83
MixUp	69.47	69.98	68.67	69.66	67.90
CutMix	75.98	74.21	69.67	72.18	68.97
FMix	72.73	70.41	68.41	68.62	66.12
GridMix	71.54	68.86	70.15	66.63	63.20
ResizeMix	69.42	68.54	67.86	63.72	63.48
SaliencyMix	69.83	69.78	70.14	68.75	67.12
PuzzleMix	73.40	73.60	70.92	71.13	69.77
AutoMix	72.91	76.24	68.44	73.40	72.10
AdAutoMix	72.83	72.63	69.66	71.43	69.69
DeiT	74.01	75.92	72.96	72.15	69.23
TransMix	75.31	76.17	74.15	72.87	71.40
SMMix	73.84	74.09	73.50	70.87	71.38
MixPro	74.78	75.26	73.49	73.18	72.28
TdAttenMix	73.63	73.32	73.11	72.19	72.12
MergeMix	**77.46**	**78.68**	**77.02**	**75.75**	**76.19**

Method	α	DeiT-S	ViT-B
Vanilla	–	86.77	91.31
MixUp	1.0	87.73	91.36
CutMix	0.2	88.37	91.53
SmoothMix	0.2	86.39	90.88
FMix	0.2	87.18	91.36
GridMix	0.2	87.58	91.31
ResizeMix	1.0	87.45	91.59
Attentive-CutMix	2.0	87.35	90.29
SaliencyMix	0.2	87.94	91.47
PuzzleMix	1.0	88.60	91.83
GuidedMix^{ap}	1.0	86.99	90.40
DeiT	0.2	88.72	92.17
TransMix	1.0	88.38	91.66
SMMix	1.0	88.76	91.93
MixPro	1.0	88.38	91.48
TdAttenMix	1.0	88.78	91.68
MergeMix	1.0	**89.42**	**92.20**

Table 3. Full system-level comparison results in LLaVA.

Models	Token[†]	Image Question Answering VQAv2	GQA	VizWiz	SciVQA[†]	TextVQA	Benchmarks MME	MMBench	MMBench^{CN}	POPE	SEED[†]	AVG	Gain
LLaVA Variants													
LLaVA-7B	Full	78.5	62.0	50.0	66.8	58.2	1510.7	64.3	58.3	85.87	66.19	65.57	–
LLaVA-NeXT-7B	Full	81.8	64.2	57.6	70.1	64.9	1519.0	67.4	60.6	86.5	70.2	69.3	–
LLaVA-NeXT-13B	Full	82.8	65.4	60.5	73.6	67.1	1575.0	70.0	64.4	86.2	71.9	71.3	–
SeVa-7B	Full	–	60.7	–	67.5	56.2	1450	65.6	59.2	86.7	65.8	–	–
SIMA	Full	–	62.2	54.4	68.1	58.3	1507.7	64.9	59.0	86.5	65.9	–	–
nSFT	Full	–	62.9	–	68.5	58.7	1531	67.1	61.0	86.8	66.2	–	–
LLaVA with Token Compressions													
LLaVA-PruMerge+	144	76.8	–	–	68.3	57.1	1462.4	64.9	–	84.0	–	–	–
VisionZip	192	77.4	60.1	–	68.2	57.8	1834.0	63.4	–	84.9	57.1	–	–
VisPrunner	128	75.8	58.2	52.7	69.1	57.0	1461.4	62.7	57.3	84.6	–	–	–
VScan	192	77.8	60.6	50.4	68.6	57.7	1806.0	63.9	57.4	86.2	–	–	–
LLaVA-Mini	1	77.6	60.9	56.2	70.4	57.0	1466.0	65.6	–	84.4	58.5	–	–
LLaVA with Augmentations & Ranking Loss													
SFT Vision	Full	**79.32**	**62.98**	47.45	70.05	57.17	**1490.88**	66.26	60.05	86.18	67.32	66.31	+0.74
+ MixUp	Full	79.27	62.58	44.95	69.41	57.39	1483.20	65.72	58.24	86.27	66.73	65.62	+0.05
+ CutMix	Full	79.18	62.40	45.04	**70.60**	57.06	1452.31	66.32	58.24	**86.47**	67.22	65.84	+0.27
+ ResizeMix	Full	77.78	61.66	44.43	68.91	55.11	1436.09	63.91	55.41	86.01	63.91	64.13	-1.44
+ MergeMix	Full	79.24	62.44	**47.69**	69.86	**57.56**	1479.97	**66.58**	**60.65**	86.10	**67.47**	**66.40**	+0.83
SFT Vision	288	78.6	**62.47**	48.15	69.51	56.41	**1486.24**	66.32	57.98	**87.37**	**66.75**	65.95	+0.38
+ MixUp	288	78.51	62.07	51.1	68.47	56.54	1459.06	65.63	**59.53**	86.86	66.06	66.08	+0.51
+ CutMix	288	78.58	62.39	50.53	**70.2**	55.95	1414.72	66.92	**59.53**	86.56	66.2	66.31	+0.74
+ ResizeMix	288	76.39	61.05	45.48	68.07	54.60	1447.35	63.31	51.97	86.57	62.54	63.33	-2.24
+ MergeMix	288	**78.61**	62.18	**52.14**	69.61	**56.85**	1453.97	66.58	59.02	86.47	66.63	**66.45**	+0.88



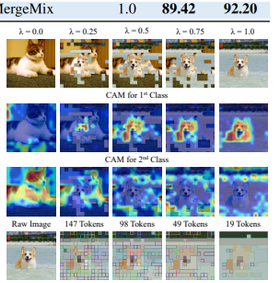


Fig 3. Visualizations.

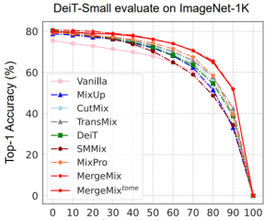


Figure 4 & 5. ECE & Occlusion.

WESTLAKE UNIVERSITY

Email: jinxin86@westlake.edu.cn

Paper: https://arxiv.org/abs/2510.23479

Web page: https://jinxins.github.io/MergeMix_Web