MM-TS: Multi-Modal Temperature and Margin Schedules for Contrastive Learning with Long-Tail Data

Supplementary Material

In the supplementary, we provide full formulation of the multi-modal contrastive InfoNCE loss, additional ablations, results on additional dataset SSv2-LT, and insights of the clusters for CC3M and YouCook2 datasets.

6. Mutli-Modal InfoNCE Loss

Given the multi-modal similarity scores:

$$s^{v\to t} = f_v(v)f_t(t)^T, \quad s^{t\to v} = f_t(t)f_v(v)^T, \tag{9}$$

The InfoNCE$(s_{t\to v})$ can be formulated as the following:

$$\mathcal{L}_{\text{InfoNCE}(s_{t\to v})} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(s_{ii}^{t\to v}/\tau)}{\sum_{j=1}^{N}\exp(s_{ij}^{t\to v}/\tau)}. \tag{10}$$

And the InfoNCE$(s_{v\to t})$ can be formulated as the following:

$$\mathcal{L}_{\text{InfoNCE}(s_{v\to t})} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(s_{ii}^{t\to v}/\tau)}{\sum_{j=1}^{N}\exp(s_{ij}^{v\to t}/\tau)}. \tag{11}$$

The the multi-modal contrastive loss is an average of the two InfoNCE losses (from vision to text and from text to vision):

$$\frac{1}{2}\left(\mathcal{L}_{\text{InfoNCE}}(s_{v\to t}) + \mathcal{L}_{\text{InfoNCE}}(s_{t\to v})\right) \tag{12}$$

7. Hyperparameters

We provide the exact values of the hyperparameters α, sh^-, and sh^+ used in the experiments. These values were selected according to the criteria outlined in the main text, ensuring positive temperatures and margins while maintaining the intended oscillation around the default values.

- **MM-TS + CLIP (CC3M):**
  $\alpha = 0.04$, $sh^- = 0.05$, $sh^+ = 0.10$
- **AVION (MI-MM loss):**
  $\alpha = 0.20$, $sh^- = 0.17$, $sh^+ = 0.30$
- **AVION (CLIP loss):**
  $\alpha = 0.08$, $sh^- = 0.05$, $sh^+ = 0.20$
- **VAST (MI-MM loss):**
  $\alpha = 0.20$, $sh^- = 0.10$, $sh^+ = 0.30$
- **VAST (CLIP loss):**
  $\alpha = 0.06$, $sh^- = 0.06$, $sh^+ = 0.09$

8. Component Analysis of MM-TS on YC2

Additionally, we provide breakdown influence of the components of our MM-TS method on YouCook2 dataset. Note, that VAST[1] has two types of evaluation with and without refinement. Without refinement evaluation is similar to the standard retrieval evaluation as with CLIP and AVION, whereas refinement includes additional training modules. In Tab. 9, we evaluate performance of VAST with InfoNCE (CLIP) and Max-Margin (MI-MM) losses. We present results with and without refinement.

9. Modality Choice for Distribution Estimation

We provide comparison training based on clusters from text embeddings (Sentence-BERT) and video embeddings (VAST backbone) on YouCook2 (see S-BERT vs VAST-v in Tab. 6). Notably, we find that video clusters also exhibit a similar long-tail distribution and modulation based on video clusters shows comparable improvement over the VAST baseline. Therefore, we conclude that while any modality can be effectively used for modulation, text is generally preferable when available.

	R@1	R@5	R@10
VAST (baseline)	50.4	74.3	80.8
S-BERT(text)	**53.0**	**77.1**	**84.5**
VAST-v(video)	52.2	76.3	84.0
CLIP-t(text)	52.5	76.4	84.4

Table 6. Comparison of retrieval performance on YouCook2 using distribution modulation based on text (S-BERT, CLIP-t) and video (VAST-v) clusters.

10. Something-Something-v2-LT dataset

Additionally, we extend our method to complex Something-Something-v2 (SSv2) [12] dataset. The dataset includes 169K training videos and 25K validation videos, each depicting interactions with everyday objects. The dataset has 174 action classes. SSv2-LT is a long-tailed subset of the original Something-Something-v2 dataset, proposed by Perrett *et al.* [38]. The resampling of training split was done using Pareto distribution with $\alpha = 6$. Validation and test splits are balanced and contain 40 and 15 samples per class respectively [38]. In [38], the authors use the standard classes for the classification task, whereas we rely on the given distribution to estimate individual shift values. Then, we train the model using full captions (without "something" placeholder) and evaluate on text and video retrieval tasks. As a base framework, we use VAST. In Tab. 7, we additionally evaluate all the components of our MM-TS framework and show the improvements with the combination of the temperature schedules and individual adjustments.

Moreover, in Tab. 8, we evaluate how robust our method to different temperatures. Our results demonstrate that the method is robust to variations of the temperature.

Method	Text-to-Video			Video-to-Text		
	R@1	R@5	R@10	R@1	R@5	R@10
VAST [1]	41.7	69.1	78.9	41.2	69.3	79.3
w/ TS	42.6	70.0	79.5	41.3	68.7	79.3
w/ TD	42.7	**70.3**	79.6	41.2	69.7	79.5
w/ TS&TD	**43.2**	**70.3**	**79.7**	**41.8**	**70.3**	**80.0**

Table 7. Performance comparison of variants of VAST on finetuning on SSv2-LT dataset. Results without refinement.

α $\tau_{\downarrow}/\tau_{\uparrow}$	0.02	0.04	0.06	0.08
0.05/0.09	42.7	43.2	43.2	42.3
0.06/0.10	**43.4**	43.0	43.1	42.8
0.07/0.11	43.0	43.2	42.9	42.8

Table 8. Impact of different temperature ranges (τ_{min} represented with $\tau_{\downarrow}$ and τ_{max} represented with $\tau_{\uparrow}$) and amplitude (α) values on the performance of VAST on SSv2-LT, evaluating R@1 Text-to-Video. The combinations marked with "-" were not done as the resulting lowest temperature for the smallest class would be ≤ 0.

11. Text Distributions

In Figs. 8, 10, we plot distributions based on 200 estimated clusters in the text embeddings for CC3M and YouCook2 datasets, respectively. In Figs. 9, 11, we plot class-based distributions for EPIC-KITCHENS-100, and SSv2-LT datasets, respectively. In Tabs. 10, 11, 12, we show random examples of captions in different clusters and manually identify the common topic between the sampled sentences.

width=!,height=!,pages=1-8

Method	Text-to-Video			Video-to-Text		
	R@1	R@5	R@10	R@1	R@5	R@10
			without refinement			
CLIP	41.4	65.0	74.6	42.6	66.8	75.4
w/ TS	43.2	67.8	77.2	44.2	70.5	79.1
w/ ICS	41.0	65.2	74.5	42.1	66.9	75.0
w/ TS&ICS	43.1	67.8	77.3	44.4	70.4	79.2
MI-MM	37.9	61.7	72.7	38.5	63.2	73.4
w/ TS	38.1	62.0	72.0	37.6	63.2	73.7
w/ ICS	36.8	60.8	71.1	37.0	62.2	72.1
w/ TS&ICS	38.1	62.8	72.7	38.1	63.2	74.2
			with refinement			
CLIP (paper)	50.4	74.3	80.8	-	-	-
CLIP *	53.1	76.1	83.2	52.6	75.9	83.5
w/ TS	53.1	77.1	84.4	52.6	76.3	84.5
w/ ICS	52.9	75.7	83.1	52.7	76.4	83.6
w/ TS&ICS	53.0	77.1	84.5	52.7	76.3	84.6
MI-MM	54.3	76.9	83.4	53.5	76.8	83.7
w/ TS	53.8	76.6	83.4	53.3	76.8	83.5
w/ ICS	54.1	76.3	82.7	54.1	77.0	83.2
w/ TS&ICS	53.9	76.7	83.7	52.8	76.8	83.7

Table 9. Performance comparison of VAST on YouCook2 dataset (Video retrieval) with modified CLIP and MI-MM losses. TS and ICS temperature (or margin) schedule and individual cluster shifts, respectively. "*"- marks our reproduction based on VAST codebase.

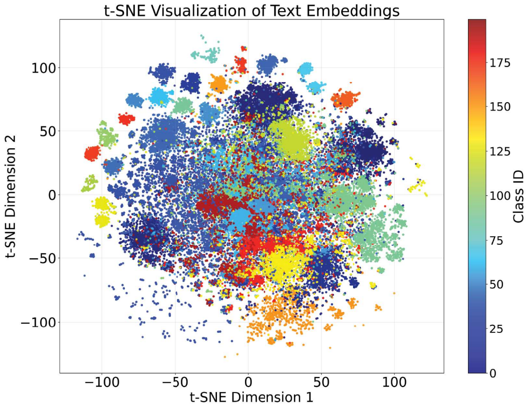

Figure 7. Visualization of the annotation embeddings in the CC3M dataset using tSNE. Each point represents a image annotation, and colors indicate the assigned clusters. In our experiments, we used 200 clusters.

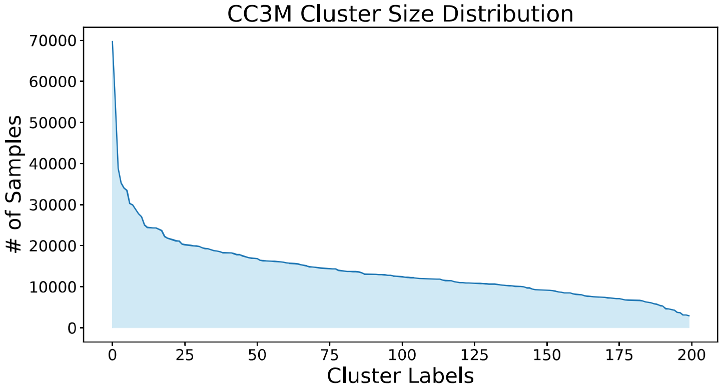

Figure 8. Visualization of long-tail annotations distribution of CC3M dataset. Annotations distribution is calculated based on k-mean clustering(200 clusters) of the annotation embeddings. Annotation embeddings are generated using BERT model [44].

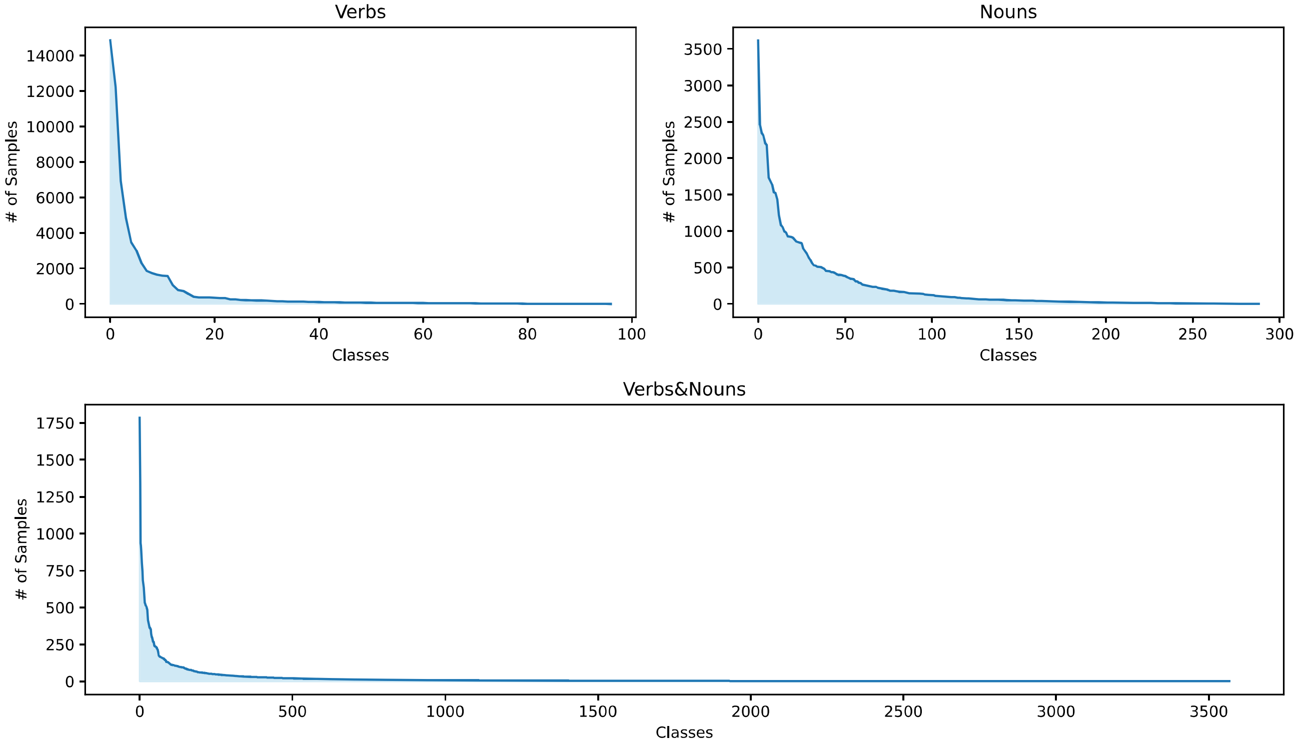

Figure 9. Visualizations of long-tail class distributions of Epic-Kitchens100 dataset training split. Class distributions are calculated based on verbs (top-left), important nouns (top-right) and unique combinations of verbs and nouns (bottom).

Cluster	Example Text Annotations	Topic
#1 106 samples	pour the egg and add a pinch of salt and red chili powder add cooking oil and beat an egg mix the eggs with salt and pepper add red pepper flakes to a bowl of eggs and whisk	eggs
#2 106 samples	add coconut milk fish sauce and soy sauce into the pan add rice vinegar soy sauce and sriracha to the pan add oil and green chilies to pan add garlic thyme bay leaf and tomato paste to the pan	pan
#3 106 samples	pour the mixture on hash browns spread the mixture loosely on the cooking sheet with olive oil and place in the oven mix the ingredients mix everything and let it cook	mixture
#4 101 samples	mix flour with the potato mixture add an egg and farmers cheese to the potatoes and mash add butter and milk to the pot and mash the potatoes add the mashed potatoes butter cream and kale to a pot and mix	potatoes
#5 92 samples	add chopped garlic chopped ginger and chopped onion add ginger garlic and onions sprinkle paprika and parsley on top mix red chili sugar garlic fish sauce and scallians with the radish	garlic
...	...	...
#196 12 samples	mix the hash brown and the sauce serve the rings with the sauce serve the beef with mashed potatoes pour gravy on the meatloaf	sauce
#197 12 samples	clean the liver and place it on a tray season the liver slice the liver place some of the liver on plastic wrap and roll it	liver
#198 11 samples	add carrot and daikon to the bowl and stir drain the carrot and daikon pickle peel and chop carrot and daikon into strips and put in a bowl rinse the daikon and carrot	carrot
#199 10 samples	scrape the hummus onto a dish and top it with some peppers and serve stir the hummus and add water blend the hummus add citric acid to the hummus	hummus
#200 9 samples	place pieces of bread into the pan and grill on both sides apply butter on one side of the bread season it with salt and pepper and put bread on grill combine 2 slices of bread and let it cook on the grill place the on the grate over the outer ring of charcoal or on the 2nd tier of a gas grill	grill

Table 10. Examples of text annotations (4 per cluster) from 5 biggest clusters (#1-#5) and 5 smallest clusters (#196-#200), obtained via kMeans clustering on SentenceBERT embeddings of training split of YouCook2 dataset.

Cluster	Example Text Annotations	Topic
#1 69669 samples	scenes of people working in the office	people
	a very long ride with the people	
	image taken from page of people	
	index : a list of the topics and people in the text	
#2 54898 samples	students are looking forward to visiting public university	students
	students have fun while they learn about the environment	
	students prepare for a perennial lesson	
	team up to help keep students safe	
#3 38840 samples	football player was praised after a sensational session between the sticks on tuesday	football player
	football player is welcomed back after winning soccer league	
	american football player will put his body on the line to help his team	
	football player is the latest player to be linked with a move away from the club	
#4 35297 samples	actor attends the european premiere of biographical film	film
	film director attends the premiere	
	film director attends premiere held at theater during festival	
	tv writer at the premiere of film	
#5 34039 samples	the truth about christmas vietnam	christmas
	merry christmas eve to all my wonderful followers ... and to everyone else as well	
	this would be cute to hang on the front door at christmas time	
	things you have to do this christmas	

Table 11. Examples of text annotations (4 per cluster) for 5 largest clusters, obtained via K-Means clustering on text embeddings generated using a distilled version of BERT on the training split of CC3M dataset.

Cluster	Example Text Annotations	Topic
#196 3736 samples	man with glasses looking himself in a mirror isolated on white background healthy fit young man measuring his waist with a tape measure to monitor his weight isolated on white background senior man and a kid passing a football isolated on white background portrait of a smiling young man isolated on white background	man on a white background
#197 3676 samples	silhouette of young boy playing cricket against a sunset background trees silhouetted against a sunset silhouette of palm trees during sunset at the beach silhouette of a dead tree with a sunset in the background	silhouette of a tree, sunset
#198 3105 samples	images from the girls basketball game images from the boys basketball game images vs. girls basketball game on thursday , february images from the girls basketball game in a city	basketball
#199 3105 samples	beautiful golden pink shining metal style uppercase or capital letter e in a 3d illustration with a shiny metallic soft purple red color classic font isolated on a white background with clipping path shiny purple metal lowercase or small letter w in a 3d illustration with a rough weathered metallic surface texture and classic font style isolated on a white background with clipping path shiny metal silver chrome beveled lowercase or small letter n in a 3d illustration with a glossy gray smooth metallic surface finish isolated on a white background with clipping path shiny gold metallic uppercase or capital letter v in a 3d illustration with a rich golden color and glossy smooth metal surface finish in a bold font isolated on a white background with clipping path	metallic letters
#200 2904 samples	sketches of cupcakes drawn by chalk on a blackboard teacher stands at the blackboard in classroom little boy to write with chalk on the school blackboard white chalk texture vintage stamp with map on a school blackboard schoolboy standing in front of a blackboard	blackboard

Table 12. Examples of text annotations (4 per cluster) for 5 smallest clusters, obtained via K-Means clustering on text embeddings generated using a distilled version of BERT on the training split of CC3M dataset.

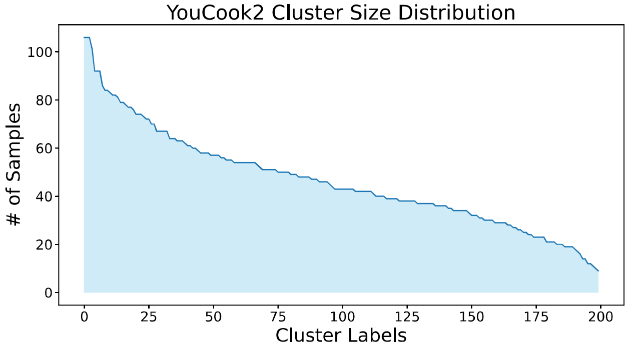

Figure 10. Visualization of long-tail annotations distribution of YouCook2 dataset. Annotations distribution is calculated based on k-mean clustering (200 clusters) of the annotation embeddings. Annotation embeddings are generated using Sentence-BERT model [43].

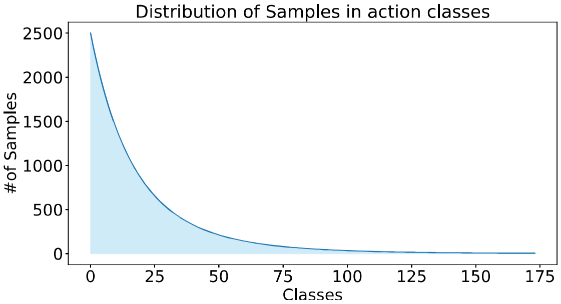

Figure 11. Visualization of action class distribution of SSv2-LT dataset.