Semantic-guided Cross-Modal Prompt Learning for Skeleton-based Zero-shot Action Recognition

Supplementary Material

In this supplementary material, we present additional information to 1) explain the remaining design choices not mentioned in the main paper; 2) support the additional ablation study on SCoPLe and list the auxiliary performance visualization not included in the main paper; 3) finalize the compatibility experiments for SCoPLe equipped with various skeleton backbones and label semantic enrichment; and 4) explore model versatility on other image-based tasks.

1. Extra Implementation Design Choices

1.1. Implementation Details of Common Prompting Baselines

In Tab. 1 of the main paper, we examine the application of various common domain universal prompting methods to address the zero-shot skeleton action recognition (ZSSAR) problem. We begin our ablation study by establishing a baseline whose text encoder is a standard **CLIP**-ViT-B/32 text encoder and consists of an MLP projection layer that maps the output features from the pre-trained skeleton backbone to the text embedding space. The subsequent baselines are established with the same projection layer configuration. Especially, **CLIP-CoOp** utilizes the prompting vectors proposed by [49] and learns to incorporate prompt-based tuning on the linguistic side. We have chosen not to include the performance evaluation of its extension method, CoCoOp, because its primary contribution is generating dynamic prompting on the image side, which is not compatible with our data inputs. **CLIP-MaPLe** references the original method in [18], which combines prefix prompting from [49] with its original layer-level prompting modules that facilitate both image and textual encoding processes. We omit its module for image processing and retain only its designs for enhancing label semantics. Finally, to provide a horizontal comparison with the text prompting module in **our model**, we remove any prompting adjustments from the skeleton branch and use only the prefix prompting along with our proposed layer-wise dual-stream language prompting module to build **CLIP-DSLP** and contrast it with the previous three baselines. The results indicate that, unlike the previous three baselines, our method maintains positive improvements compared to the results of directly using CLIP.

1.2. Design Choices of δ for Generalized Zero-shot Skeleton Action Recognition

We used a simple gating mechanism to implement generalized ZSSAR in our solution, establishing an entropy threshold of δ for binary classification based on the logits calculated from $\mathcal{L}_{test}(V_y^x, F_y)$. Since each testing split protocol has different ratios and distributions of seen and unseen classes, we found that using a constant δ across all experiments often fails. Following [14], we set aside a few samples from the original NTU dataset to build a gating validation set for each split arrangement and conducted an individual hyperparameter grid search on δ. We selected the best hyperparameter (see Tab. 1) for each experiment and used it for the final performance evaluation.

Testing Protocol			δ
General	NTU-60	55/5 split	4.007331371
		48/12 split	3.871198177
	NTU-120	110/10 split	4.700477600
		96/24 split	4.564345360
Random Split	NTU-60	55/5 split	4.007330418
	NTU-120	110/10 split	4.700477600
	PKU-MMD	46/5 split	3.828638792

Table 1. Hyperparameter selection of δ for each testing protocol in the General Performance experiments and the Random Split experiments.

2. Extra Experiment Results

2.1. Visualized Unseen Class Accuracy

Following the operations in SA-DVAE [22], we conducted a similar per-class accuracy visualization and compared it to the results in [22] using the same testing protocol in the random split experiment on the 55/5 split of NTU-60. The previous paper demonstrated a significant improvement in achieving preliminary discrimination between classes that share very similar visual appearances at a whole-skeleton scale. Specifically, the unseen class "wear a shoe" has a very similar visual appearance and textual constitution to the seen class "take off a shoe", while also being semantically surrounded by other interconnected actions like "standing up". This makes naive alignment easily lean towards classifying labels 9 and 16 into the same category, as they both share similar visual patterns learned from the seen class "take off a shoe". [22] mitigated this issue by strictly disentangling the label-related skeleton features for more robust and refined cross-modal alignment synergy. In our model, we used dual-stream language prompting to effectively inherit the original semantic generalization ability of the language model and interacted it with joint-level prompts to reach a similar effect in skeleton feature tuning as [22] for refined cross-modal alignment. The results in Fig. 1 show that not only could

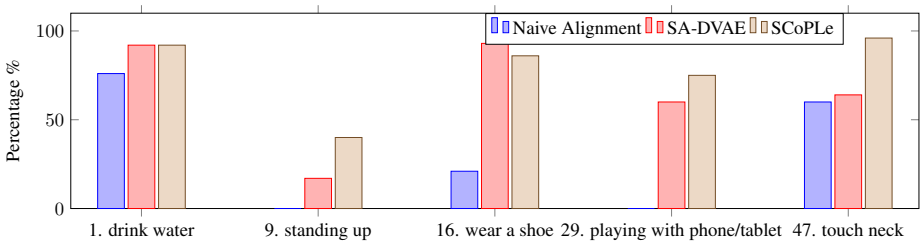

Figure 1. Unseen per-class accuracy of the 55/5 split testing protocol on the NTU-60 dataset. The unseen split $\{1, 9, 16, 29, 47\}$ is used in a challenging run of the random-split GZSL experiments.

our method compete with the prediction accuracy of class 16 from [22], but we also improved the overall model performance on other unseen classes.

2.2. Layer-level Dual-stream Fusion

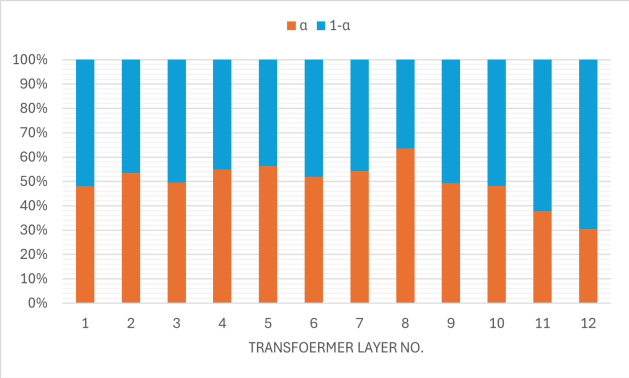

Figure 2. An example illustration of learned α and $1-\alpha$ belonging to each Transformer layer of the CLIP text encoder for SCoPLe trained on the 110/10 testing split of NTU-120.

In our dual-stream language prompting module, the original CLIP text encoder receives two streams of label text encoding flows: the original and prompted encoding streams. To maintain the interaction between the two streams, we preserve a learnable layer-level α throughout each prompted Transformer layer. Fig. 2 provides an illustrative example of all learned α values for each layer of a pre-trained CLIP text encoder for SCoPLe, under the experiment protocol of the 110/10 testing split on NTU-120. According to the weighted aggregation in Eq. (5) in the main paper, α controls the contribution weights of the prompted encoding contexts, while $1-\alpha$ controls the weights of the original encoding contexts. The results show that α_s vary across layers, with significant changes beyond the eighth layer (0.3 to 0.63), indicating the different importance assigned to prompt tuning. We find that for some layers, they indeed benefit from the original semantics and demonstrate gradient learning that increases the attention guidance from the original encoding stream.

2.3. Visualization of Adaptive Visual Representation Sampler

To ensure cross-modal prompting synergy, we design the visual-side prompting to be compatible with a joint-level feature format and utilize an Adaptive Visual Representation Sampler to apply text-semantic-guided weighted sampling on real joint features and skeleton prompts. Fig. 3 provides an example visualization of the sampling weights predicted by a trained SCoPLe against the prompted skeleton feature of an incoming example belonging to a given seen/unseen class in the testing protocol of a 110/10 split on NTU-120. Based on the visualization, we can come up with two observations. For seen class classification, the method ensures high prediction precision and effectively resolves challenging recognition tasks, such as distinguishing between "reading" and "writing", by learning to focus on the visual features related to the joints located in various positions for the hands. Since the training stage emphasizes direct pattern learning for seen classes, prompting has a minor contextual influence and primarily functions as feature tuning for adapting cross-modal knowledge transfer. For unseen classes, the model can still allocate reasonable visual focus to collect semantically relevant skeleton feature cues by relying on the cross-modal contextual supplements from the visual prompts. For example, for "punch other person", SCoPLe spends balanced attention to the joints around the spine and the tip of the left hand. For "drop", SCoPLe concentrates attention on the joints around the left thumb.

2.4. Extra ZSL Split Study

Here we provide more results for skeleton-based ZSL using more splitting protocols from [51]. Considering the utilization rate of the original CLIP semantic knowledge during the evaluation, we compared our method against the baseline results of using only the CLIP encoder for cross-modal alignment without any feature refinement from the previous paper. According to Tab. 2, as the proportions of the unseen classes increased, the original labels encountered transfer limitations, causing both models to experience decreased accuracies. However, with the support of our prompting modules, SCoPLe can mitigate the deterioration to a certain

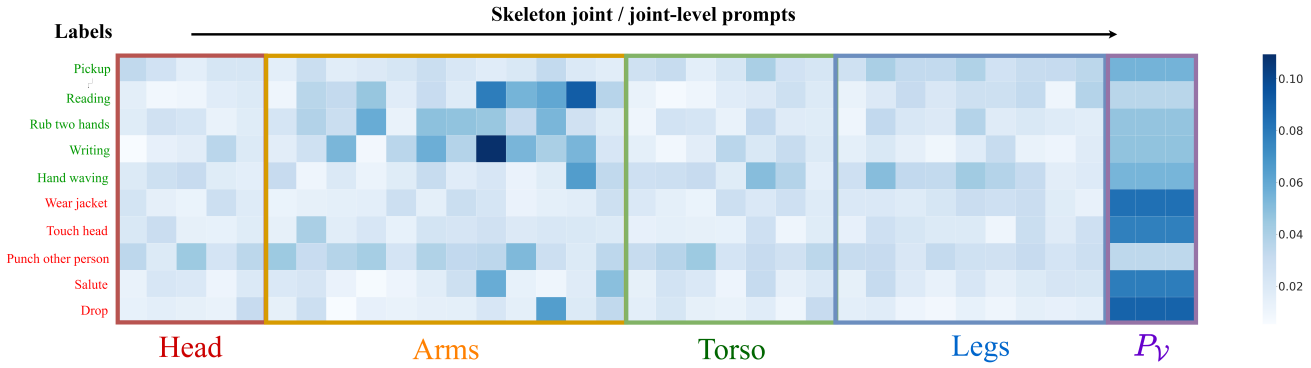

Figure 3. An illustrative example of semantic-guided sampling weights for each skeleton joint and joint-level prompt in a given action sample. One row provides the sampling weights for all joints/prompts of a data sample that belongs to the action class labeled on the left. Each cell represents a single joint feature or prompt vector for the corresponding sample, while $P_{\mathcal{V}}$ is shared across all inputs. Green action samples represent instances of seen classes, while red action samples represent instances of unseen classes. Each joint is manually assigned to a known body part for clarity.

extent.

	NTU-RGB+D 60				
	55/5 split	48/12 split	40/20 split	30/30 split	20/40 split
CLIP	62.58	33.16	27.15	16.29	7.23
SCoPLe	84.10	52.96	32.00	18.17	8.46
	NTU-RGB+D 120				
	110/10 split	96/24 split	80/40 split	60/60 split	40/80 split
CLIP	39.13	48.15	21.31	14.12	5.24
SCoPLe	63.51	52.17	25.31	15.67	7.39

Table 2. ZSL metrics (%) for PURLS [51] and SCoPLe on the additional split experiments from [51].

2.5. Extra Ablation Study

Tab. 3 presents our extra ablation study conducted on our method, combining each module according to different assembly strategies. This includes using non-dual-stream language prompting with the full skeleton prompting module (first row) and using skeleton prompting without the sampler and the full DSLP (second row). The resting ablation study has already been fully explored in the main paper. We use the full version of SCoPLe as the contrast baseline and carry out the experiments using the testing protocols for general performance evaluation on NTU-60 and NTU-120. Consistent with the conclusions we reached in Sec.4.4. in the main paper, incorporating partial prompting mechanisms provides only limited and unstable improvements in prediction.

3. Combination with Existing Methods

3.1. Skeleton Extractors

Our method is a powerful, effective, and flexible plug-in framework that is compatible with various skeleton backbones to conduct cross-modal alignment with CLIP, provided that the encoder output is in the format of joint-shaped features. In Tab. 4, we follow [51] and use different modern skeleton backbones to verify the compatibility of SCoPLe. We show that SCoPLe can efficiently raise the upper limit of ZSSAR performance for each skeleton backbone with semantic assistance from CLIP.

3.2. Label Semantic Enrichment

In [50] and [22], the authors propose extension experiments that utilize a large language model (such as ChatGPT) to augment class descriptions with richer action-related information. We report results using the same settings for both the general performance and the random split testing protocols. As shown in the last two rows of Tab. 5, our prompting method can still adapt to the new semantic environment and further improve the results obtained with the original labels. Similarly, in the horizontal performance comparison with the results in [22], although the degree of optimization on ZSSAR is not as pronounced as in SA-DVAE, our method benefits more from the label semantic enrichment in improving GZSSAR results.

4. Versatility of DSLP

Our proposed tuning designs of DSLP focus on enhancing CLIP's generalization to unseen classes by reducing prompt tuning biases from seen-class training. While the primary goal is to address a ZSSAR test case as outlined in the main paper, we believe that our feature-level approach can also theoretically benefit tasks such as image-based base-to-novel, few-shot learning, and other similar scenarios when novel class discrimination is involved. We applied our text prompting approach (DSLP) to the linguistic branch of MaPLe [18] and conducted some classic base-to-novel generalization experiments recorded in [18] on OxfordPets, Caltech101,

Model	NTU-60								NTU-120							
	s55/5 split				s48/12 split				s110/10 split				s96/24 split			
	ZSL	S	U	H	ZSL	S	U	H	ZSL	S	U	H	ZSL	S	U	H
SCoPLe w. text prompt (no dual) + skel. prompt (full)	70.20	59.01	56.26	57.60	39.60	36.31	79.73	49.89	62.34	51.21	48.35	49.74	47.01	56.87	45.12	50.30
SCoPLe w. text prompt (full) + skel. prompt (no sampler)	71.78	77.27	53.93	63.52	42.97	54.03	59.09	56.45	69.65	64.62	55.35	59.62	50.73	52.38	50.96	51.66
SCoPLe (full)	**84.10**	69.60	71.94	**70.75**	**52.96**	54.49	61.83	**57.93**	**74.53**	63.51	61.08	**62.27**	**52.17**	53.33	51.18	**52.23**

Table 3. The extra ablation analysis for selectively combining each module in SCoPLe across all testing splits for ZSL and GZSL on the NTU datasets. Other ablation results are already listed in the main paper.

Skeleton Backbone	NTU-60							
	s55/5 split				s48/12 split			
	ZSL	S	U	H	ZSL	S	U	H
AA-GCN	51.07	28.00	84.44	42.06	24.89	38.71	74.73	51.00
AA-GCN (w. SCoPLe)	61.69	64.48	63.65	64.06	31.61	61.29	59.36	60.31
CTR-GCN	72.48	29.52	76.89	42.67	30.96	71.21	40.45	51.59
CTR-GCN (w. SCoPLe)	74.75	76.37	69.18	72.60	36.71	54.45	68.99	60.86
DG-GCN	68.47	60.20	61.11	60.65	36.99	72.23	42.12	53.21
DG-GCN (w. SCoPLe)	76.99	82.23	62.12	70.77	42.92	59.08	64.79	61.81
Shift-GCN	62.58	41.16	80.63	54.50	33.16	53.78	58.35	55.97
Shift-GCN (w. SCoPLe)	84.10	69.60	71.94	70.75	52.96	54.49	61.83	57.93
Shift-GCN + augmented text	70.88	65.64	56.77	60.88	41.03	61.41	53.10	56.96
Shift-GCN (w. SCoPLe) + augmented text	82.52	81.72	68.69	74.64	54.20	56.81	61.91	59.25

Table 4. ZSL & GZSL metrics (%) for the CLIP baseline and SCoPLe equipped with different skeleton backbones and LLM-augmented class descriptions on the general performance testing protocol of the NTU-60 dataset.

Method	NTU-60 55/5 split				NTU-120 110/10 split			
	ZSL	S	U	H	ZSL	S	U	H
SMIE [50]	65.08	/	/	/	46.4	/	/	/
SMIE + augmented text	70.89	/	/	/	52.04	/	/	/
SA-DVAE [22]	84.2	78.16	72.6	75.27	50.67	58.09	40.23	47.54
SA-DVAE + augmented text	87.61	74.54	76.5	75.51	57.16	53.32	48.36	50.72
SCoPLe	83.72	75.32	80.17	77.67	53.34	70.47	44.29	54.08
SCoPLe + augmented text	84.34	77.38	75.30	76.32	56.73	77.89	45.25	57.24

Table 5. ZSL & GZSL metrics (%) for SMIE, SA-DVAE and SCoPLe with LLM-augmented class descriptions on the random split testing protocols of the NTU-60 and NTU-120 datasets.

EuroSAT, DTD, and FGVCAircraft to evaluate the versatility of our module. As shown in Tab. 6, our method can achieve the most consistent harmonic prediction means in these scenarios and demonstrate greater robustness in both base and novel prediction advantages in most cases.

	OxfordPets			Caltech101			EuroSAT		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP	91.17	97.26	94.12	96.84	94.00	95.40	56.48	64.05	60.03
CoOp	93.67	95.29	94.47	98.00	89.81	93.73	92.19	54.74	68.69
Co-CoOp	95.20	97.69	96.43	97.96	93.81	95.84	87.49	60.04	71.21
MaPLe	95.43	97.76	96.58	97.74	94.36	96.02	94.07	73.23	82.35
Ours (DSLP)	**95.85**	**98.48**	**97.15**	**98.12**	**95.10**	**96.58**	**95.16**	**73.98**	**83.24**
	DTD			FGVCAircraft					
	Base	Novel	HM	Base	Novel	HM			
CLIP	53.24	59.90	56.37	27.19	36.29	31.09			
CoOp	79.44	41.18	54.24	**40.44**	22.30	28.75			
Co-CoOp	77.01	56.00	64.85	33.41	23.71	27.74			
MaPLe	80.36	59.18	68.16	37.44	35.61	36.50			
Ours (DSLP)	**79.59**	**63.04**	**70.35**	37.53	**36.86**	**37.19**			

Table 6. Comparison with CLIP [30], CoOp [49], Co-CoOp [48] and MaPLe [18] base-to-novel generalization for image recognition.