A APPENDIX

The appendix is structured as follows:
§A.1 provides experimental results on weakly-supervised image-text matching with pseudo captions generated by GIT, CoCa, and BLIP-2.
§A.2 demonstrates the expression format of pseudo captions generated by the image captioning tools.
§A.3 provides the detailed derivation of anchor selection.
§A.4 provides the detailed derivation of Equation 9.
§A.5 provides experimental results on text-based person retrieval.
§A.6 provides experimental results on noisy correspondence learning.
§A.7 provides experimental results on zero-shot image classification of CLIP pre-training.
§A.8 provides experimental results on VLP fine-tuning.
§A.9 provides efficiency analysis of AdaCL.
§A.10 provides a comprehensive study of other hyper-parameters.
§A.11 provides an analysis on additional arguments involved in anchor selection.
§A.12 provides a more comprehensive visualization results of clone negatives.
§A.13 provides a discussion of the limitations for this work.

A.1 WEAKLY-SUPERVISED IMAGE-TEXT MATCHING

In the main manuscript, we utilize BLIP(Li et al., 2022) to generate pseudo captions based on Flickr30K training set. Then, the original text annotations are replaced by the pseudo captions for training to verify the robustness of AdaCL in handling clone negatives. To mitigate concerns about reliance on a specific captioning tool, we conduct a range of complementary experiments, to comprehensively analyze its robustness. Specifically, four captioning tools are selected, i.e., GIT(Wang et al., 2022a), CoCa(Yu et al., 2022), and BLIP-2(Li et al., 2023). GIT is a multi-modal pre-training method that unifies vision-language tasks such as image/video captioning and question answering. CoCa employs a unified transformer architecture to perform both image-text matching and image captioning tasks. CoCa is trained on large-scale image-text pairs and can generate descriptive captions for images while also understanding the relationship between visual and textual content. BLIP-2 is an advanced vision-language model that builds upon its predecessor, BLIP. It introduces a lightweight Querying Transformer (Q-Former) to bridge pre-trained vision and language models efficiently.

The way of generating pseudo captions is unified, i.e., through the zero-shot image captioning results. The maximum length of the pseudo caption is 30. Table 1 demonstrates the matching results of different baselines with AdaCL based on pseudo captions. The experimental settings are the same with the settings in the main manuscript.

From Table 1, we can make the following conclusions: By employing different image captioning method, AdaCL demonstrates matching performance that are within a **3%** margin. AdaCL achieves highly competitive performance on the four annotation settings. Therefore, AdaCL is further proved to be applicable to more general annotation settings, where the issue of clone negatives is well mitigated. The impact of image captioning methods on AdaCL is trivial. More specifically, AdaCL-$\mathcal{X}$ (BLIP) and AdaCL-$\mathcal{X}$ (BLIP-2) performs slightly better than AdaCL-$\mathcal{X}$ (GIT) and AdaCL-$\mathcal{X}$ (CoCa). We speculate that this is because the pseudo captions generated by GIT and CoCa typically focus on action and instance information of the image gallery, while the pseudo captions generated by BLIP possess general descriptions, often consisting of a subject-verb-object structure with more holistic-level semantics. Overall, the retrieval performance is quite comparable, further verifying the robustness of AdaCL.

A.2 EXPRESSION FORMAT OF PSEUDO CAPTIONS

Regarding pseudo captions, we have presented a subset of caption cases generated by four distinct captioning tools (BLIP, GIT, BLIP-2, and CoCa), as illustrated in Figure 1. Pseudo captions generally provide a global and coarse overview of the images, encompassing more potential clone negatives. Among the four methods, GIT produces the most concise captions, while CoCa tends to generate more detailed descriptions. The weakly-supervised image-text matching based on pseudo captions

Table 1: Comparisons of Image-Text Retrieval performance on Flickr30K test set with pseudo captions generated by four distinct captioning methods. AdaCL-$\mathcal{X}$ (BLIP), AdaCL-$\mathcal{X}$ (GIT), AdaCL-$\mathcal{X}$ (CoCa), and AdaCL-$\mathcal{X}$ (BLIP-2) represent AdaCL with respective image captioning methods. **Bold** is the best performance, while red indicates the margin between the best and worst.

Methods	Image→Text			Text→Image		
	R@1	R@5	R@10	R@1	R@5	R@10
AdaCL-CMPM (BLIP)	**46.3**	72.9	**85.8**	34.1	63.6	75.0
AdaCL-CMPM (GIT)	44.5 (-1.8)	71.2	84.9 (-0.9)	32.7 (-1.8)	61.0 (-2.7)	73.7 (-1.5)
AdaCL-CMPM (CoCa)	44.8	71.0 (-2.4)	85.1	34.1	**63.7**	74.6
AdaCL-CMPM (BLIP-2)	45.9	**73.4**	85.7	**34.5**	63.3	**75.2**
AdaCL-SCAN (BLIP)	59.2	86.9	94.7	41.7	**73.2**	84.1
AdaCL-SCAN (GIT)	**60.0**	**87.7**	**94.9**	41.1	70.9 (-2.3)	83.4 (-1.6)
AdaCL-SCAN (CoCa)	58.1 (-1.9)	85.2	94.2 (-0.7)	40.2 (-2.0)	72.6	83.6
AdaCL-SCAN (BLIP-2)	58.1	85.2 (-2.5)	94.6	**42.2**	74.1	**85.0**
AdaCL-CVSE (BLIP)	64.7	82.6	**92.9**	**47.0**	77.5	88.4
AdaCL-CVSE (GIT)	63.6 (-1.9)	80.9 (-1.8)	90.9 (-2.0)	46.2	77.4	**88.7**
AdaCL-CVSE (CoCa)	63.8	81.4	91.5	45.3 (-1.7)	77.1 (-0.6)	88.0 (-0.7)
AdaCL-CVSE (BLIP-2)	**65.5**	**82.7**	92.3	46.3	**77.7**	88.2
AdaCL-DIME (BLIP)	**71.3**	**88.3**	**94.9**	54.7	82.8	90.4
AdaCL-DIME (GIT)	70.1 (-1.2)	88.0	94.2	55.1	82.3	90.6
AdaCL-DIME (CoCa)	70.8	87.8 (-0.5)	93.7	54.6 (-1.3)	82.0 (-1.4)	90.0 (-1.0)
AdaCL-DIME (BLIP-2)	70.4	**88.3**	93.7 (-1.2)	**55.9**	**83.4**	**91.0**

poses imposes demands for handling clone negatives, thus rendering this task more challenging. We will release all datasets based on pseudo captions to facilitate further research in this domain.

A.3 Derivation of m_1 and m_2 in AdaCL

Revisiting AdaCL, **our goal is to progressively tune m_1 and m_2 based on the *anchor*.** Therefore, we first copy Equation 3 in the main manuscript, i.e., softmax normalized similarity for each image I and its corresponding text T with two margin parameters, which can be expressed as:

$$\hat{p}_i(I) = \frac{\exp\left[m_1(s(I,T_i) - m_2)\right]}{\exp\left[m_1(s(I,T_i) - m_2)\right] + \sum_{j-1,j\neq i}^{M+1} \exp\left[s(I,T_j)\right]}, \tag{1}$$

$$\mathcal{L}_{ada} = \mathbb{E}_{I\sim D}\left[\mathbb{H}(\mathbf{y}(I), \mathbf{p}(I))\right] = -\frac{1}{N}\sum_{i=1}^{N} y_i(I)\log(\hat{p}_i(I)). \tag{2}$$

The potential in-batch clone negatives are represented as:$S^* := \{s \mid p\left(\mathcal{C} \mid s\right) > p\left(\bar{\mathcal{C}} \mid s\right)\}$, and *anchor* is defined as the median of S^*. The two specific boundary functions of the anchor are defined as:

$$\hat{p}_u = \frac{\exp\left[m_1(\textit{anchor} - m_2)\right]}{\exp\left[m_1(\textit{anchor} - m_2)\right] + \sum_{\textit{anchor}}}, \tag{3}$$

$$\frac{\exp\left[m_1(1 - m_2)\right]}{\exp\left[m_1(1 - m_2)\right] + \sum_{\textit{anchor}}} = 1 - \epsilon, \tag{4}$$

where $\sum_{\textit{anchor}}$ is the simplification of $\sum_{k=1,k\neq u}^{M+1} \exp\left[s(I_u,T_k)\right]$. $\sum_{\textit{anchor}}$ can be obtained through Equation 3, expressed as:

$$\sum_{\textit{anchor}} = \frac{1-\hat{p}_u}{\hat{p}_u} e^{m_1(\textit{anchor}-m_2)}. \tag{5}$$

Combining Equation 5 and Equation 4, we have:

$$\left[e^{m_1(1-m_2)} + \frac{1-\hat{p}_u}{\hat{p}_u} e^{m_1(\textit{anchor}-m_2)}\right](1-\epsilon) = e^{m_1(1-m_2)}. \tag{6}$$

Image Query	Ground-truth	BLIP	GIT	BLIP-2	CoCa
	A band is playing to a cheering concert with many people.	A crowd of people in a concert with a band on stage.	Many people are enjoying a concert.	A crowd of people in a large concert.	A large crowd of people are gathered in a concert.
	Music being played by several individuals while a crowd sits and listens.	People sitting in chairs watching a man play a musical instrument.	People having an indoor concert.	A group of people and a man playing a violin	A group of musicians sitting in a room with instruments .
	Two men who are riding on a horse both are trying to rope a bull in a rodeo.	There is a man riding a horse with a cow.	A man riding a horse chasing the cow.	A man is on a horse	A man on a horse roping a calf in a rodeo.
	A man wearing blue jeans is trying to stop a horse.	A man is falling off of a horse.	A man is riding a horse.	A man is trying to catch a horse that is running away.	A man in a black and white striped shirt is trying to rope a horse .
	A crowded sidewalk in the inner city of an Asian country.	People walking on a street in a city with shops.	People walking on the street.	A group of people walking down a street	A group of people walking down a street .
	A crowd of people is walking down the middle of a city street.	There are many people walking down the street together.	Many people gather on the street.	A large crowd of people walking down a street.	A crowd of people walking down a street .
	A crowd of people in running outfits runs a marathon with two skyscrapers in the background.	People running in a marathon in a city with tall buildings.	People are jogging during the day.	A group of people running in a city	A group of people walking on the sidewalk near a building .
	A group of people is running a race or marathon in the city.	People are running in a marathon in a city street	Athletes running in the city.	A large crowd of people running in a marathon	A group of people that are standing in the street .
	A man in a blue T-shirt speaks into a blow horn towards a group of people.	People are walking in a crowded city street.	Cyclists riding across the street.	A man is standing on a street corner with a megaphone	A group of people riding bikes down a street .
	Many people are chilling in front an old building.	People are in front of a large building with a clock tower.	Tourists sitting at tables outside the building.	A crowd of people sitting at tables outside of a building.	A crowd of people sitting at tables in front of a building.
	A group of people stand in the park of a city, with buildings in the background.	People standing around a fountain in a city with tall buildings	Several people are standing in the park.	A group of people standing in the park.	A group of people standing in front of a water fountain .

Figure 1: Cases of pseudo captions by four distinct captioning tools, i.e., BLIP, GIT, BLIP-2, and CoCa.

To simplify Equation 6, we have:

$$\frac{(1-\epsilon)(1-\hat{p}_u)}{\hat{p}_u} e^{m_1 \cdot anchor} = \epsilon \cdot e^{m_1}. \tag{7}$$

By taking the logarithm of both sides of Equation 7, we have:

$$m_1 \cdot anchor + log \frac{(1-\epsilon)(1-\hat{p}_u)}{\hat{p}_u} = m_1 + log\ \epsilon. \tag{8}$$

Then, m_1 can be obtained, expressed as:

$$m_1 = log(\frac{\epsilon\ \hat{p}_u}{(1-\epsilon)(1-\hat{p}_u)})/(anchor - 1), \tag{9}$$

which corresponds to Algorithm 1 in the manuscript. Meanwhile, by taking the logarithm of both sides of Equation 5, we have:

$$log \sum\nolimits_{anchor} = log \frac{1-\hat{p}_u}{\hat{p}_u} + m_1(anchor - m_2). \tag{10}$$

Simplifying Equation 10, we obtain m_2:

$$m_2 = anchor + log(\frac{1-\hat{p}_u}{\hat{p}_u \cdot \sum\nolimits_{anchor}})/m_1, \tag{11}$$

which corresponds to Algorithm 1 in the main manuscript. With Equation 9 and Equation 11, m_1 and m_2 can be computed and updated during each batch training process with the supervision of *anchor*, facilitating the model to exploit more distinguishable cross-modal semantics among samples compared with the original TRL and CL.

A.4 Derivation of Equation 9

Here we demonstrate the derivation of Equation 9. To begin with, a K-class classification probability with Bayes's formula can be expressed as:

$$p(y=i \mid x) = \frac{p(x \mid y=i)p(y=i)}{\sum_{j=1}^{K} p(x \mid y=j)p(y=j)} = \frac{\exp\left(f_i(x)\right)}{\sum_{j=1}^{K} \exp\left(f_j(x)\right)}, \tag{12}$$

In anchor selection of AdaCL, the input variable x (a.k.a s) is one-dimensional with a binary output variable $y \in 0, 1$ (a.k.a $\bar{\mathcal{C}}$ and $\mathcal{C}$). We aim to predict $p(y=1 \mid x)$. Since GDA assumes that for each class $y=0$ and $y=1$, the input x follows a gaussian distribution. This can be expressed as: $p(x \mid y=0) = \mathcal{N}\left(x \mid \mu_0, \sigma_0^2\right)$ and $p(x \mid y=1) = \mathcal{N}\left(x \mid \mu_1, \sigma_1^2\right)$. μ_0, μ_1 and σ_0^2, σ_1^2 are the means and variances of distributions for classes $y=0$ and $y=1$, respectively. Thus, the posterior probability can be expressed as:

$$p(y=0 \mid x) = \frac{\mathcal{N}\left(x \mid \mu_0, \sigma_0^2\right) \cdot \pi_0}{\mathcal{N}\left(x \mid \mu_0, \sigma_0^2\right) \cdot \pi_0 + \mathcal{N}\left(x \mid \mu_1, \sigma_1^2\right) \cdot \pi_1}, \tag{13}$$

$$p(y=1 \mid x) = \frac{\mathcal{N}\left(x \mid \mu_1, \sigma_1^2\right) \cdot \pi_1}{\mathcal{N}\left(x \mid \mu_1, \sigma_1^2\right) \cdot \pi_1 + \mathcal{N}\left(x \mid \mu_0, \sigma_0^2\right) \cdot \pi_0}. \tag{14}$$

Since $\mathcal{N}\left(x \mid \mu_1, \sigma_1^2\right)$ is the probability density function of a Gaussian distribution, we substitute $\mathcal{N}\left(x \mid \mu_1, \sigma_1^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ into the above equations, obtaining Equation 9 in the manuscript:

$$p\left(y=0 \mid x\right) = \frac{1}{1 + \frac{\pi_1}{\pi_0}\frac{\sigma_0}{\sigma_1} \exp\left[\frac{(s-\mu_0)^2}{2\sigma_0^2} - \frac{(s-\mu_1)^2}{2\sigma_1^2}\right]}, \tag{15}$$

$$p\left(y=1 \mid x\right) = \frac{1}{1 + \frac{\pi_0}{\pi_1}\frac{\sigma_1}{\sigma_0} \exp\left[\frac{(s-\mu_1)^2}{2\sigma_1^2} - \frac{(s-\mu_0)^2}{2\sigma_0^2}\right]}, \tag{16}$$

$p\left(y=0 \mid x\right)$ and $p\left(y=1 \mid x\right)$ represent the probability of a similarity score to be a clone negative or not, without the need of explicit pre-processing to the dataset or training.

Table 2: R@1 Results on text-based person search. "DG" stands for domain generalization, and "FT" for fine-tuning on the corresponding dataset.

Method	CUHK-PEDES		ICFG-PEDES		RSTPReid	
	DG	FT	DG	FT	DG	FT
CL	16.3	49.3	15.8	43.5	12.7	30.1
AdaCL	**30.5**	**56.7**	**27.9**	**49.0**	**23.9**	**41.4**

A.5 Domain Generalization on Text-based Person Retrieval

To evaluate the robustness of AdaCL as a plug-and-play module, we seek to evaluate its domain generalization capabilities in text-based person retrieval. Specifically, we conduct training using Flickr30K and select three mainstream text-based person retrieval datasets, CUHK-PEDES(Li et al., 2017), ICFG-PEDES(Ding et al., 2021), and RSTPReid(Zhu et al., 2021) for domain generalization experiments. To ensure a fair comparison, we choose CMPM as the baseline, as the learning objective it adopt in its paper is the closest to vanilla contrastive learning, and its original paper indeed conducted experiments on two of the datasets. As illustrated in Table 2, it is observed that AdaCL boosts CMPM by a large margin. Especially for DG, the results of AdaCL on each dataset improves by over 10%.

Table 3: Fine-tuning results of AdaCL on three baselines under ICFG-PEDES.

Method	Text-Image R@1	Text-Image R@5	Text-Image R@10
CMPM	43.5	65.4	74.2
AdaCL-CMPM	**49.0**	**69.7**	**79.1**
ViTAA	51.0	68.8	75.8
AdaCL-ViTAA	**54.8**	**74.1**	**78.6**
IRRA	63.5	80.3	85.8
AdaCL-IRRA	**64.3**	**81.1**	**86.5**

Table 4: Fine-tuning results of AdaCL on three baselines under RSTPReid.

Method	Text-Image R@1	Text-Image R@5	Text-Image R@10
CMPM	30.1	38.5	59.6
AdaCL-CMPM	**41.4**	**57.0**	**55.7**
ViTAA	37.7	60.6	66.5
AdaCL-ViTAA	**42.6**	**62.1**	**69.2**
IRRA	60.2	81.3	88.2
AdaCL-IRRA	**62.7**	**81.4**	**89.0**

For fine-tuning, in addition to CUHK-PEDES, we also validate the performance of AdaCL on ICFG-PEDES and RSTPReid. Three baselines are employed: CMPM(Zhang & Lu, 2018), ViTAA(Wang et al., 2020), and IRRA(Jiang & Ye, 2023), and compare the effectiveness of incorporating AdaCL as a constraint. The experimental results w/ and w/o using AdaCL are presented in Table 3 and Table 4. It is observed that AdaCL demonstrates significant improvements across the three baselines, achieving absolute enhancements of 5.5%, 3.8%, and 0.8% in R@1, respectively. These matching results substantiate the robustness of AdaCL in other vision-language downstream tasks, demonstrating its insensitivity to the diverse dataset distributions (both natural images and person search images), and the choice of baselines.

A.6 Noisy Correspondence Learning

As mentioned in Seection 1, noisy correspondence learning (NC) (Huang et al., 2021; Yang et al., 2023; Ma et al., 2024; Qin et al., 2023) focuses on handling negatives by manually introducing noisy labels. Several works classify samples into clean and noisy subsets, followed by a rectifier and triplet

ranking loss to boost the learning of NC. We further validate AdaCL in such challenging scenarios by plugging in AdaCL and verify its NC effectiveness on Flickr30K using the same pre-processing strategy (by shuffling the captions of training images for a specific percentage, denoted by noise ratio). The matching results under two noise ratio (20% and 40%) are reported in Table 5.

Table 5: Noisy correspondence learning of AdaCL. We follow (Huang et al., 2021) to shuffle the captions of training images for a specific percentage, i.e., noise ratio.

Noise Ratio	Methods	Image→Text			Text→Image		
		R@1	R@5	R@10	R@1	R@5	R@10
20%	NCR	75.0	93.9	97.5	58.3	83.0	89.0
	AdaCL-NCR	75.3	93.8	97.4	61.2	84.1	89.7
	BiCro	78.1	94.4	97.5	60.4	84.4	89.9
	AdaCL-BiCro	79.6	95.2	97.5	62.7	85.1	91.3
	CREAM	77.4	95.0	97.3	58.7	84.1	89.8
	AdaCL-CREAM	80.0	95.6	97.4	61.9	86.4	91.3
	CRCL	77.9	95.4	98.3	60.9	84.7	90.6
	AdaCL-CRCL	81.0	96.2	98.5	62.3	84.9	91.7
40%	NCR	68.1	89.6	94.8	51.4	78.4	84.8
	AdaCL-NCR	74.7	92.3	96.6	57.8	82.0	87.1
	BiCro	74.6	92.7	96.2	55.5	81.1	87.4
	AdaCL-BiCro	75.3	93.1	96.2	57.4	82.5	89.6
	CREAM	76.3	93.4	97.1	57.0	82.6	88.7
	AdaCL-CREAM	79.2	95.1	98.3	61.5	86.0	90.2
	CRCL	77.8	95.2	98.0	60.0	84.0	90.2
	AdaCL-CRCL	80.3	95.0	98.1	61.7	84.4	90.9

We also validate the effectiveness of AdaCL on CC152K. CC152K consists of 150,000 samples from training split of Conceptual Captions (CC) (Sharma et al., 2018) for training, 1,000 samples from validation split for validation, and 1,000 samples from validation split for testing. As all image-text pairs in CC are automatically harvested from the Internet, approximately 3%–20% of the pairs in the dataset are mismatched or weakly matched. This benchmark aligns well with the settings of NC, making it a suitable choice for evaluating AdaCL.

From Table 5 and Table 6, it can be concluded that for a noise ratio of 20%, AdaCL achieves notable improvements, particularly in I-T R@1 (AdaCL-CRCL improves from 77.9 to 81.0) and T-I R@1 (AdaCL-NCR improves from 58.3 to 61.2). For a noise ratio of 40%, the trend of improvement remains consistent, although the performance naturally decreases as noise increases. Notably, AdaCL-CRCL demonstrates strong robustness with I-T R@1 improving from 77.8 to 80.3, even at high noise levels. While the baseline results degrade significantly as the noise ratio increases, AdaCL exhibits better resilience, as evidenced in AdaCL-NCR (I-T R@1 only drops from 75.3 to 74.7). AdaCL's robustness is particularly evident in T-I matching, where the decline in performance is less pronounced compared to the baselines (AdaCL-CRCL achieves T-I R@5 of 84.4 at 40% noise ratio). Similar to Flickr30K, AdaCL also demonstrates consistent improvements over the baselines on CC152K. The performance improvements of AdaCL on both datasets further support its generalizability and applicability in noisy correspondence learning.

Table 6: Noisy correspondence learning of AdaCL on CC152K.

Methods	Image→Text			Text→Image		
	R@1	R@5	R@10	R@1	R@5	R@10
NCR	39.5	64.5	73.5	40.3	64.6	73.2
AdaCL-NCR	43.2	66.9	74.9	42.5	69.0	76.2
BiCro	40.8	67.2	76.1	42.1	67.6	76.4
AdaCL-BiCro	42.9	66.1	76.0	42.7	68.4	78.7
CREAM	40.3	68.5	77.1	40.2	68.2	78.3
AdaCL-CREAM	43.1	69.6	77.2	42.2	70.0	80.2
CRCL	41.8	67.4	76.5	41.6	68.0	78.4
AdaCL-CRCL	42.4	68.0	77.4	41.7	69.3	80.0

A.7 Zero-shot Image Classification of AdaCL in CLIP pre-training

In addition to image-text matching, we also evaluate AdaCL on other pre-training task, i.e., zero-shot image classification. Specifically, we validate AdaCL on eight common classification benchmarks, which can be divided into (i) general datasets: ImageNet(Deng et al., 2009), CIFAR-10(Krizhevsky et al., 2009), CIFAR-100(Krizhevsky et al., 2009), Caltech-101(Fei-Fei et al., 2004)), and (ii) fine-grained datasets: Food-101(Bossard et al., 2014), Flowers-102(Nilsback & Zisserman, 2008), OxfordPets(Parkhi et al., 2012), and FGVCAircraft(Maji et al., 2013). The Top-1 accuracy results of "CLIP + AdaCL" pretrained on CC3M and CC12M are demonstrated in Table 7:

Table 7: Zero-shot image classification of CLIP pre-training under different learning objectives. "Baseline" represents "CLIP+vanilla contrastive learning", and "AdaCL" represents "CLIP+AdaCL". Results under two pre-training settings, i.e., CC3M and CC12M are compared.

Data	**Model**	**Datasets**							
		ImageNet	CIFAR-10	CIFAR-100	Caltech-101	Food-101	Flowers	Pets	Aircraft
CC3M	Baseline	17.2	71.3	32.1	50.9	10.2	10.8	12.1	1.0
	AdaCL	**22.0**	**77.1**	**42.2**	**54.8**	**12.6**	**13.3**	**14.9**	**1.7**
CC12M	Baseline	32.9	72.5	38.0	74.0	26.5	**25.7**	46.2	2.6
	AdaCL	**34.8**	**73.4**	**43.3**	**74.7**	**33.1**	25.4	**46.7**	**2.8**

It is observed that AdaCL outperforms CL in all the general datasets and most of the fine-grained datasets, proving its advantage in recognition tasks. Specifically, in ImageNet, CIFAR-10, CIFAR-100, the Top-1 accuracy of AdaCL has surpassed vanilla CL by over 5%. It is noteworthy that the performance on fine-grained datasets further verifies AdaCL's capacity in challenging scenarios.

A.8 VLP Fine-tuning

Apart from CLIP pre-training, we further report the fine-tuning results of AdaCL in several Vision Language Pre-training methods (VLP) by fine-tuning them using AdaCL on MS-COCO (5K). As illustrated in Table 8, AdaCL facilitates matching performance across nearly all metrics under both dual-encoder method (BEIT-3(Wang et al., 2022b)) and fusion-encoder methods (UNITER(Chen et al., 2020), OSCAR(Li et al., 2020), VinVL(Zhang et al., 2021)), effectively boosting the fine-tuning process. These results further corroborate the robustness of AdaCL across multiple baselines.

Table 8: Results of AdaCL on VLP fine-tuning.

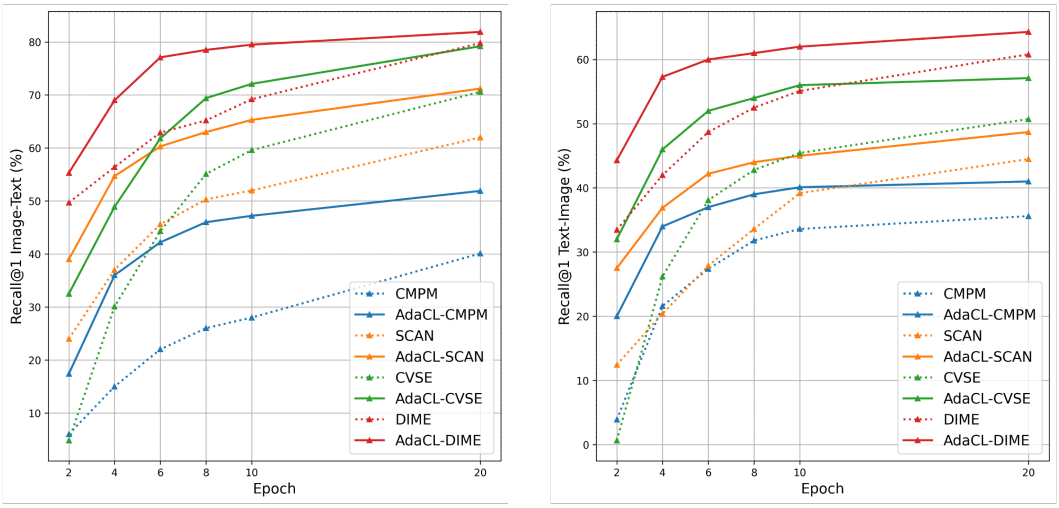
Methods	Image→Text			Text→Image		
	R@1	R@5	R@10	R@1	R@5	R@10
UNITER†	65.7	88.6	93.8	52.9	79.9	88.0
AdaCL-UNITER	**67.6**	**89.0**	**94.3**	**55.1**	**81.2**	**88.9**
OSCAR	70.0	91.1	95.5	54.0	<u>80.8</u>	88.5
AdaCL-OSCAR	**71.0**	**92.7**	**96.3**	**54.0**	80.6	**89.1**
VinVL	75.4	92.9	96.2	58.8	83.5	90.3
AdaCL-VinVL	**78.7**	**94.4**	**96.8**	**60.4**	**84.2**	**91.1**
BEIT-3	<u>84.8</u>	96.5	98.3	67.2	87.7	92.8
AdaCL-BEIT-3	84.4	**96.9**	98.3	**68.6**	**89.1**	**93.7**

† Evaluated by us with official repository.

A.9 Efficiency Analysis

Serving as a plug-and-play module, AdaCL does not increase the inference time since it is independent of the cross-modal reasoning module. For training efficiency, we add detailed analysis on AdaCL.

The speed of model convergence mirrors the learning efficiency of a certain constraint. As shown in Figure 2, AdaCL brings considerable convergence efficiency and retrieval results, which even boost CMPM and DIME to achieve their ultimate results within the first 5 epochs, demonstrating the scalability and efficiency of AdaCL.



(a) Convergence efficiency of Image-text retrieval

(b) Convergence efficiency of Text-image retrieval

Figure 2: Training efficiency of AdaCL.

A.10 Ablation studies of other hyper-parameters

A.10.1 Analysis of momentum memory bank

Since memory bank is widely adopted in vision-language contrastive learning, we further analyze AdaCL by varying the memory bank size M, which leads to different number of negative samples. The results in Table 9 reveal that the momentum memory bank yields a modest yet discernible improvement: Among the sizes of 4096, 6144, and 8192, the impact of memory bank is not that significant. This suggests that AdaCL does not excessively rely on the quantity of negative samples for an ideal similarity distribution.

Table 9: Effect of different memory bank sizes.

Memory Bank Size	Image→Text			Text→Image		
	R@1	R@5	R@10	R@1	R@5	R@10
N/A	70.3	90.0	95.5	49.5	77.5	87.3
2048	71.9	90.9	96.7	51.4	78.4	87.3
4096	**74.2**	91.7	97.9	**53.7**	81.1	**88.2**
6144	73.7	**91.9**	**98.2**	53.2	79.6	87.8
8192	73.9	91.4	98.0	53.3	**80.5**	87.8

A.10.2 Analysis of mini-batch size

Since mini-batch size is correlated with the number of potential anchor candidates for selection, we also investigate the impact of mini-batch size, as shown in Table 10. It can be observed that AdaCL exhibits remarkable robustness to variations in batch size settings. Across a range of batch sizes from 16 to 128, the fluctuation in R@1 remains within a narrow margin of 4%.

Table 10: Effect of different mini-batch sizes.

Batch Size	Image→Text			Text→Image		
	R@1	R@5	R@10	R@1	R@5	R@10
16	70.6	89.2	96.5	50.0	79.3	84.8
32	72.9	90.9	97.4	52.1	81.0	87.1
64	**74.2**	**91.7**	**97.9**	53.7	81.1	**88.2**
128	74.0	91.2	97.6	**53.9**	**81.2**	88.0

A.11 Additional parameters in anchor selection

Here we further analyze the anchor selection methodology. In the main manuscript, we utilize the negatives from S_{sln} and S_{cln} within each mini-batch to represent the empirical means and variances. However, we cannot guarantee that all samples in S_{sln} and S_{cln} are exclusively salient negatives and clone negatives. Based on this speculation, we continue to select Top-K similarity scores from S_{sln} and S_{cln} as observational samples for calculating means and variances. The assumption of this study is higher similarity scores for certain negatives correlate with an increased probability of them being clone negatives. We set K to 32 and conduct experiments on CMPM, SCAN, and DIME under Flickr30K, as shown in Table 11. It can be concluded that employing Top-K selection strategy does not result in a significant improvement or deterioration in matching performance, with fluctuations generally remaining within a 2% range. This observation contradicts our initial hypothesis and intuition. Consequently, we can infer that AdaCL exhibits ***low sensitivity to the specific values of the empirical mean and variance***, which is another minor merit. Given that the Top-K selection explicitly increase computation without yielding significant performance improvements, we have opted to maintain the original calculation method in the main manuscript.

Table 11: Matching results of Top-K selection for empirical means and variances.

Methods	Image→Text			Text→Image		
	R@1	R@5	R@10	R@1	R@5	R@10
AdaCL-CMPM	54.7	79.0	87.5	41.6	69.4	79.2
AdaCL-CMPM †	54.2	77.8	87.1	41.8	68.1	80.0
AdaCL-SCAN	71.4	93.0	97.2	50.9	79.9	86.8
AdaCL-SCAN †	72.7	93.4	96.5	50.2	79.0	87.1
AdaCL-DIME	82.6	96.3	98.9	63.6	88.4	93.7
AdaCL-DIME †	82.4	95.3	98.7	63.7	88.2	93.0

†: Employ Top-K selection.

A.12 More Visualization of AdaCL

We present a more comprehensive comparison of CL, TRL, and AdaCL trained with ground-truth annotations and pseudo captions. The visualization results of the early training stage are demonstrated in Figure 3, which include 4 kinds of clone negatives with 11 cases. Based on the attention maps, we can summarize the following conclusions: AdaCL captures abundant semantics on highly similar clone negatives. Specifically, case (a) and case (b) demonstrate that AdaCL boosts the exploration of spatial semantics among the images, such as "music being played by several individuals", as well as "is trying to stop a horse", which effectively distinguishes clone negatives apart. Additionally, case (c) demonstrates AdaCL's ability in capturing background information such as "Asian country" and "a city street" are crucial phrases that are reasoned through AdaCL. Case (d) showcases five examples of urban landscape, demonstrating that AdaCL is able to discover instances that are not explicitly described in the text query. For instance, the unique attribute "spectator" is not included in Q8, but AdaCL facilitate learning the corresponding representation, which is highlighted in the attention map. Also, the latent "fountain" is not included in Q11 but reasoned by AdaCL. In this way, AdaCL is proved to achieve comprehensive cross-modal semantics with its adaptive tuning strategy even when

the quality of textual annotations is not high. This finding presents great potential of AdaCL to handle retrieval with low quality labels.

Furthermore, we obtain the attention maps by training with pseudo captions under AdaCL, as depicted in the last column of Figure 3. Due to the lack of instance-level information during the training process, we do not expect the results to surpass models trained on original annotations. However, AdaCL (Pseudo Caption) manages to capture the approximate cross-modal semantics and pays attention to the fine-grained representation, which outperforms CL and TRL (trained with ground-truth) in most cases. This demonstrates the prospects of AdaCL in the vision-language contrastive learning of automatically annotated image-text pairs.

A.13 Discussion: Limitation

In this work, AdaCL is evaluated on (1) image-text matching under Flickr30K, MS-COCO, (2) CLIP pre-training under CC3M and CC12M, (3) weakly-supervised image-text matching under pseudo captions, (4) text-based person search under CUHK-PEDES, ICFG-PEDES, and RSTPReid. We have not extended AdaCL to an all-round vision-language tasks due to time and computational limitations, which is undoubtedly planned in our future endeavor.

Also, although AdaCL maintains high convergence efficiency, we acknowledge that AdaCL inevitably introduces additional computation during training with a moderate computational overhead of $O(N \cdot M)$ per batch training. We believe this trade-off is acceptable given the context of contrastive learning and pre-training. In future work, we will delve into a more lightweight vision-language learning paradigm.

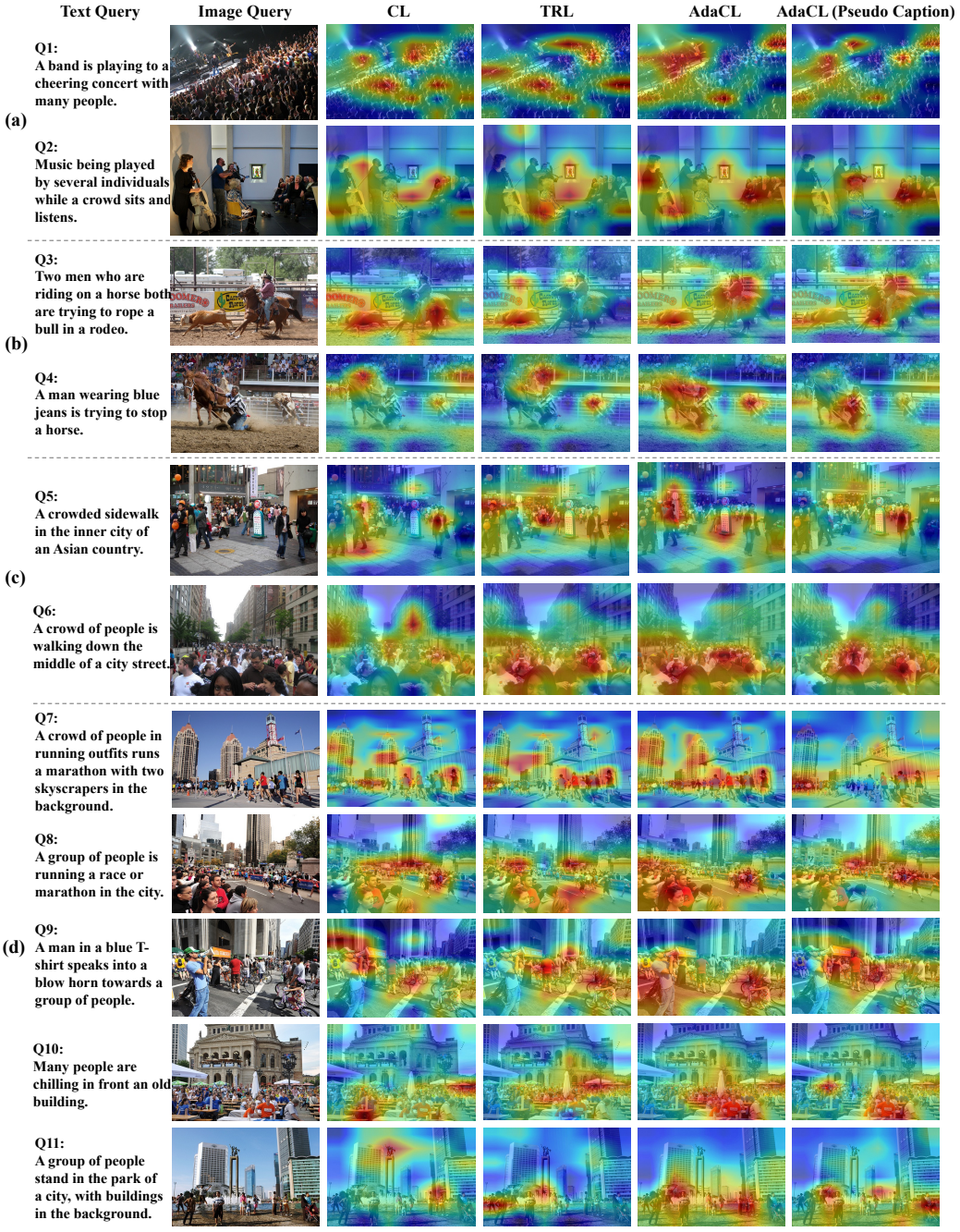


Figure 3: Attention maps of clone negative cases in early stage (Epoch 10). "CL", "TRL", and "AdaCL" represent model trained with different constraints. The last column represents AdaCL trained with pseudo captions.

REFERENCES

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, pp. 446–461. Springer, 2014.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pp. 104–120. Springer, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Zefeng Ding, Changxing Ding, Zhiyin Shao, and Dacheng Tao. Semantically self-aligned network for text-to-image part-aware person re-identification. *arXiv preprint arXiv:2107.12666*, 2021.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pp. 178–178. IEEE, 2004.

Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. Learning with noisy correspondence for cross-modal matching. *Advances in Neural Information Processing Systems*, 34:29406–29419, 2021.

Ding Jiang and Mang Ye. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1970–1979, 2017.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pp. 121–137. Springer, 2020.

Xinran Ma, Mouxing Yang, Yunfan Li, Peng Hu, Jiancheng Lv, and Xi Peng. Cross-modal retrieval with noisy correspondence via consistency refining and mining. *IEEE Transactions on Image Processing*, 2024.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pp. 722–729. IEEE, 2008.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

Yang Qin, Yuan Sun, Dezhong Peng, Joey Tianyi Zhou, Xi Peng, and Peng Hu. Cross-modal active complementary learning with self-refining correspondence. *Advances in Neural Information Processing Systems*, 36:24829–24840, 2023.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022a.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022b.

Zhe Wang, Zhiyuan Fang, Jun Wang, and Yezhou Yang. Vitaa: Visual-textual attributes alignment in person search by natural language. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pp. 402–420. Springer, 2020.

Shuo Yang, Zhaopan Xu, Kai Wang, Yang You, Hongxun Yao, Tongliang Liu, and Min Xu. Bicro: Noisy correspondence rectification for multi-modality data via bi-directional cross-modal similarity consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19883–19892, 2023.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5579–5588, 2021.

Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 686–701, 2018.

Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua. Dssl: Deep surroundings-person separation learning for text-based person retrieval. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 209–217, 2021.