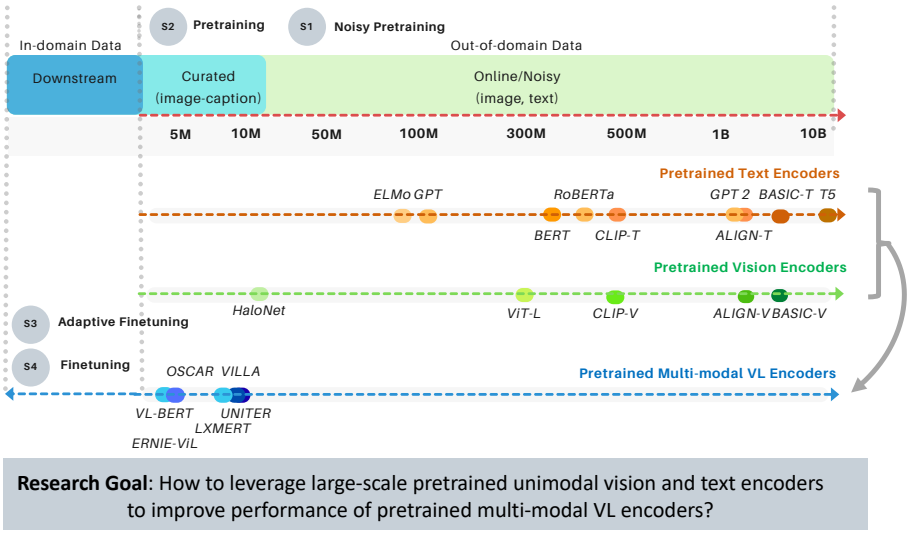

Figure 1: Comparison of data size between different pretraining frameworks. It is noticeable that the data size of pretraining unimodal encoder models is generally much larger than Vision-Language models.

A-DUE: Adaptive Distillation Leveraging Pretrained Unimodal Encoders for Vision-Language Tasks

Anonymous authors
Paper under double-blind review

Abstract

Prior work demonstrates the importance of leveraging knowledge in pre-trained encoders for Vision-Language (VL) tasks. Therefore, several studies focus on developing a better framework to fuse vision and language information in the pre-training stage. In this paper, we take a different route and investigate how to distill knowledge from various vision, language, and vision-language encoders to fine-tune a strong downstream model. Our method, A-DUE, avoids redoing any pretraining step and only requires access to downstream data. Moreover, it is also generalizable to be applied on top of different VL models. To comprehensively demonstrate the effectiveness of A-DUE, we conduct evaluation across three main VL benchmarks *i.e.* VCR, VQA, and SNLI-VE under four settings *i.e.* zero-shot, low-shot, full-shot and domain-shifted settings. Results prove that A-DUE consistently improves the performance with different VL base models across all the settings. It even achieves state-of-the-art on VCR among models pretrained with image-text data and delivers competitive results on VQA and SNLI-VE. Via our analysis, we also discover that A-DUE can assist VL models in becoming more robust in predictions and regulate them in better utilizing visual information. Code will be made available.

1 INTRODUCTION

Existing approaches pretrained Vision-Language (VL) encoders on multimodal data such as image captioning datasets achieve impressive performance. However, curated image caption and other forms of vision-language aligned data are label extensive and even impossible for some domains. Recently, other works **?** have explored utilizing largely crawled noisy image and text pairs of data for pre-training. However, comparatively it's always easier to obtain large (unpaired) unimodal data comparing to paired cross-modal data considering the fact that the existing largest cross-modal data is less than xxM, but there are several unimodal data set exceeding xxB in size **?**. Those unimodal data requires significantly less manual labor and are collected from many different domains. Given the existence of text and vision encoders pretrained on large unimodal data across various domains, it raises a natural question whether we can leverage these largely pretrained unimodal encoders to improve a vision-language model on downstream tasks without the need to access their original data.

As larger and larger unimodal encoders appear in the field, this becomes an important question to be further resolved. In this paper we investigate to answer this question via proposing our Multimodal Adaptive Distillation (MAD) framework. We further justify the contribution of MAD by evaluating it on top of different base models across multiple VL benchmarks, especially the highly-semantic ones like Visual Commonsense Reasoning (VCR) **?** and Visual Entailment (SNLI-VE) **?**.

Conventional methods naively plug in the pretrained unimodal encoders into a downstream VL structure and redo the whole pre-training step (typically in 10M scale of pre-training data) which leads to tremendous computation expense. In reality, larger and larger pretrained unimodal encoders come out every year and it is extremely inefficient if there is a need to redo the pre-training step every time.

In contrast, our framework is flexible that it does not require access to any of the pre-training data and can combine information from multiple pretrained unimodal encoders to benefit the VL models with training on the task-specific data only. The pretrained vision and text encoders would serve as the dual teachers and adaptively distill their knowledge to the VL base model, the student. This distillation framework is generalizable as it can be applied on top of different VL base models. Additionally, the various combination of selecting different unimodal encoders as the teacher set enables other researchers to flexibly discover and optimize to discover the teacher set for other downstream tasks.

For thoroughly demonstrating the supremacy of our framework, we evaluate our framework on top of different base models across multiple VL tasks *i.e.* (VCR, SNLI-VE, and VQA) and further under a variety of data availability constraints, including zero-shot, low-shot, and fully-supervised settings. For VCR, we additionally use established ways of perturbing the characteristics of the evaluation set to measure model performance under domain-shift. From our comprehensive evaluation, we illustrate that applying MAD is able to assist the base VL models consistently improve their performance, achieve the State-Of-The-Art (SOTA) on VCR Q2A for single models pretrained with image-text data, and competitive results on SNLI-VE and VQA. We even discover interesting finding that the existing VL models tend to under-utilize the visual information and applying MAD on top of them is able to boost the utilization.

2 RELATED WORK

Vision and Language Pretraining: Since the success of text pretraining with attention-based models, such as BERT **?**, GPT **?**, etc., those pretrained text encoder models have been broadly employed in different tasks. VL models, such as LXMERT **?**, VL-BERT **?**, UNITER **?**, VILLA **?**, and others **??????** followed to integrate the pretrained text models into their frameworks. Thus they have to conduct additional VL pretraining (VLP) with image-caption data. All of these pretraining steps utilize additional datasets on the order of 10M samples for various pretraining objectives, including Masked Language Modeling (MLM), Image-Text Matching (ITM), *etc.* Similarly large-scale pretrained vision models can be obtained via vision pretraining frameworks such as ViT **?** on image dataset including JFT-300M **?**. Trying to utilize large-scale pretrained vision models, a recent work, CLIP-ViL experiments to **?** adapt the vision encoder from CLIP. Similarly, this also unavoidably results in redoing all the VLP steps. Both pretrained vision and text models can also be obtained via contrastive pretraining frameworks which reserve modality-independence **???**. These

contrastive learning frameworks can also be regarded as a shallow method to fuse pretrained vision and text encoders for VL tasks. They heavily rely on cosine distance between output features of text and vision encoder models. However, without incorporating pretrained vision, text models into a pretrained VL model but a shallow cosine measure, these frameworks have not proved to produce a high performance on highly semantic VL tasks like VCR, *etc.*

Currently, there is a lack of a more general framework for fusing large-scale pretrained vision and text models into a pretrained VL structure for improving downstream VL tasks while controlling computational complexity. Existing fusion methods, such as shallow contrastive frameworks, that avoid impacting computation complexity mostly focus on traditional VL tasks such as object classification, image captioning, *etc.* To our best knowledge, no prior VLP works from this tier have studied the impact on generalization capability by assessing performance on highly-semantic tasks, such as VCR, especially under both low-shot and domain shifted scenarios, which are more reflective of the challenges encountered in practice.

In this work, we address several of these mentioned gaps. First, we propose an efficient distillation approach to leverage pretrained vision and text encoder models into a pretrained VL structure in a way that doesn't require additional pretraining, and retains inference complexity. Second, we study the impact of utilizing large-scale pretrained vision, text encoders like CLIP-V, CLIP-T, Roberta, ViT, *etc.* on VL tasks including VCR, SNLI-VE, and VQA under true zero-shot, low-shot, and domain shifted scenarios.

Generalization of Question Answering: Visual question answering **??** is among the most challenging VL tasks, due to large variation between samples and distribution discrepancy between training and testing datasets. This results in difficulty for models to perform well under zero-shot, low-shot, and domain-shifted settings. Despite high accuracy recent models achieve **?????**, other works have begun to uncover the tendency of models to leverage spurious shortcut signals, or memorization of mapping distributions **????**. While prior works have begun to explore question answering tasks under zero-shot **??**, low-shot **??** and domain shifted **???????** settings, they are mostly limited within text-only question answering and low-level visual question answering tasks (*i.e.* VQA **?**). To the best of our knowledge, none of the prior works have explored true zero-shot and low-shot settings in highly complex visual question answering datasets such as VCR or SNLI-VE, and only one prior work **?** contributed to VCR with domain shift.

In this work, we conduct a thorough evaluation with many existing top-performing models on zero-shot, low-shot, and domain-shifted settings.

Knowledge Distillation: In conventional knowledge distillation **????**, a larger model serves as the teacher to a smaller student model. The goal is usually to obtain a computationally lighter and more efficient framework but still maintain similar or even higher accuracy **??????**. However, recently many frameworks with small model complexity may have large pretraining data which results in more generalized feature representations. Crucial values still exist for distillation from those large-scale pretrained models to downstream domain-specific models with potentially larger model complexity. It is shallow to only compare the inference complexity between them ignoring the large pretraining complexity.

Also, tremendous progress was made in knowledge distillation with unimodal data. For instance, in vision, **??** propose distillation for visual representation learning. In addition, remarkable advances have been made in knowledge distillation for language model compression **???**. A wide range of works have explored to supervise the student model by mimicking different components of the teacher, *e.g.* the distribution of self-attention, intermediate representations of transformer blocks, last layer features, *etc.* to increase performances **???**. On the other hand, only one prior explored knowledge distillation in VL transformers **?** while it mainly follows the conventional methods with unimodal distillation. The distillation is not differentiated between vision and language. Further, in most of these scenarios, the feature components leveraged for the distillation loss is fixed, and the weight of the distillation loss is also fixed. Prior works related to attention structures only focus on fixed sequence-to-sequence distillation **?** and only one prior work first introduced the idea of dynamically adjusting the distillation loss weight on a per-instance level **?**.

In this work, our framework, for the first time, proposes multimodal distillation which modulates the distillation between modalities enabling flexible switch of different teacher models for each modality. Our approach is also model-agnostic ignoring the inference complexity comparison between the

teacher and the student. The experiments prove that it is effective in all of these scenarios. We also leverage the idea of dynamic distillation weights, but we even go one step further: we dynamically adjust not only the *distillation weight* for each sample instance but *which token features are distilled* as well within every sequence.

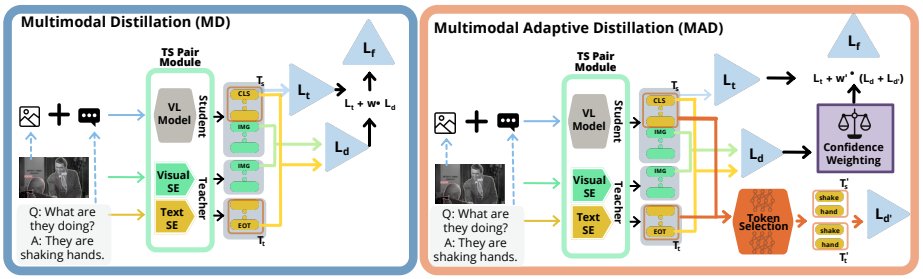

Figure 2: Structure diagram of Multimodal Distillation (MD) and Multimodal Adaptive Distillation (MAD). MAD is further improved on top of MD and MD is equivalent to MAD in terms of structure when both Confidence Weighting and Token Selection are removed.

3 METHOD

In this work, we utilize pretrained knowledge from large-scale pretrained unimodal encoders to help downstream VL tasks. We achieve this via separately distilling knowledge from the visual teacher model and the text teacher model into the base student VL model without further re-accessing the pre-training data and increasing the computation complexity in inference time of the VL model. We introduce Multimodal Aaptive Distillation (MAD) framework in the following order: First, Multimodal Distillation (MD) in transformers; Secondly, distillation with Confidence Weighting (CW); Then, distillation with Text Token Selection (TTS); Lastly, two-stage adaptive finetuning.

3.1 MULTIMODAL DISTILLATION

Distillation in Transformer: Most of the existing largely pretrained unimodal encoders and high-performing downstream VL models are transformer-based. For distillation from a transformer-based model to another, the conventional method is to extract the special [CLS] tokens (typically the first token in a sequence) from both transformers **????** and measure the token-level feature difference. Hence, the distilled knowledge is not differentiable between modalities.

Multimodal Distillation in Transformer: In this work, we aim to separate the distillation process for visual and text information. So we can flexibly have different combinations of a visual and a text encoder as dual teacher models from different pre-training.

To achieve this, in the teacher-level, depending on the type of text encoders, we would either select the [CLS] or [EOS] token as the representative token [1]. Further we utilize the token feature to represent the information of the whole text sequence. In the student-level, we also select the [CLS] token's feature to represent the whole text sequence. As for the visual information, in the teacher-level, we follow the design of the existing largely pretrained transformer-based visual encoders to also utilize the feature of the leading classification token to represent the whole visual sequence. Further, considering each element in the visual sequence is the extracted feature of an ROI from the input image, in the student-level, we purposely create an aggregated mean-pooled visual feature across all the visual tokens to represent the whole image. We then inject the feature as a special token, [IMG] to the beginning of the visual sequence to represent the whole sequence.

After extracting the representative visual and text tokens' features from both the teacher and student models, we calculate the token-level feature difference for the visual and text tokens separately via L1 measure to obtain the corresponding distillation loss.

$$L_{d,t} = \left\| f_{t-t}\left(t\right) - f_{s-t}\left(t\right) \right\|_1 \tag{1}$$

[1] *e.g.* [EOS] for CLIP-T and [CLS] for RoBERTa

where t refers to the input text; f_{t-t} and f_{s-t} refer to the feature extraction of the representative text tokens from the teacher and student models respectively.

$$L_{d,v} = \|f_{t-v}(i) - f_{s-v}(\partial(i))\|_1 \tag{2}$$

where i refers to the input image; f_{t-v} and f_{s-v} refer to the feature extraction of the representative visual tokens from the teacher and student models respectively.

After summing the distillation losses from both modalities, we multiply the sum with a distillation weight w. When conducting multimodal distillation, the final loss is consisted of both the downstream task loss $L_{downstream-task}$ and the distillation loss. Formally,

$$L_{downstream-task} + w \cdot (L_{d,v} + L_{d,t}) \tag{3}$$

3.2 Distillation with Confidence Weighting

The data distribution in downstream benchmarks may include domain-specific (in-domain) distribution. For instance, the questions in VCR mostly focus on human intentions, emotions and temporal information and the images are all from movies. Despite the pretrained knowledge of unimodal encoders are learned across multiple domains and helpful in obtaining generalized VL representations, the pretrained data is relatively out-of-domain distribution. Hence, it is difficult to guarantee if the pretrained knowledge is worth distilled for every sample in the downstream finetuning. Considering this, we design an approach to toggle the distillation weight depending on the ratio between confidence scores of the teacher and student models.

$$w = \frac{1}{T} \cdot S\left(\frac{\max(\sigma(L^t))}{\max(\sigma(L^s))}\right) \tag{4}$$

where L^t represents the logit vector from the teacher model and L^s from the base student model. σ represents the sofmax function and S represents the sigmoid function. T is a learnable weight that initializes at 10. When the ratio between confidences is above 1, the teacher model is more confident with its prediction, thus the distillation weight w would be increased accordingly. Otherwise, the distillation weight would be reduced to prevent the interference of the teacher model.

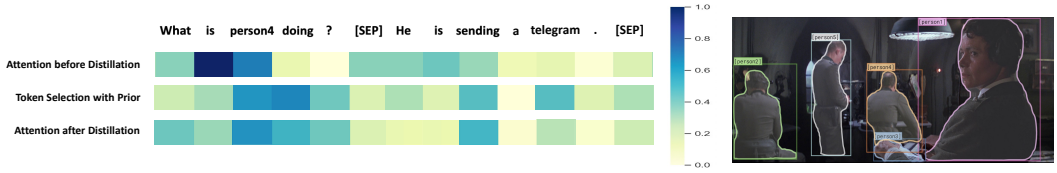

Figure 3: Comparison of attentions from VL-BERT before and after distillation, and Token Selective Module scores.

3.3 Distillation with Text Token Selection

The conventional methods **????** distill knowledge from the teacher transformer to the student transformer in a sequence-to-sequence level. However, former works **?** point out, for an input text sequence, existing models still struggle in differentiating semantically important text tokens from the trivial ones. As an example in Fig. 3, among the input text sequence, tokens like "person4", "doing", "sending" and "telegram" are semantically important since they contain crucial information for understanding the visual scene and answering the question. Yet, existing models may mistakenly focus on trivial tokens like "is" and "[SEP]" and neglect the important tokens. This may further result in failure to really understand the visual scene and picking up shortcuts. To resolve this, we intend to incorporate language priors to help differentiate semantically important tokens from others in every input text sequence. Consequently, we can further utilize the selected tokens to guide the knowledge distillation procedure in a more refined token-level. This may help the student model in learning the semantic information of the selected tokens instead of shortcuts or biases from the trivial ones.

Given a text sequence $t = \{w_0 \dots w_z\}$, where z is the length of the sequence, we apply a pretrained Token Selection Module (TSM) $f_{tsm}(i, t)$ to discriminate the semantically important tokens. TSM generates a score, s_l for each token $w_l, l \in [0, z]$. and every s_l is a summation of two subsequent scores, $s_{l,v}$ and $s_{l,t}$.

We first apply a pretrained CLIP model **?** to measure the visual relevance between every token and the input image. Thereby, $S_v = \{cos \langle f_{c-v}(i), f_{c-t}(w_l) \rangle, l \in [0, z]\}$. Here, f_{c-v} represents the feature extraction by CLIP's visual encoder and f_{c-t} for the text encoder. The cosine distance between the extracted visual and text features would be served as the visual relevance score.

We further adapt a pretrained keyword extraction model **?** to produce the syntactic importance score for each token, $s_{l,t}$

After summation, $S = \frac{S_v}{|S_v|_1} + \frac{S_t}{|S_t|_1}$ We then rank text tokens based on S and select m tokens with the highest scores, t', where $|t'| = m$. The corresponding features of both teacher and student model would be compared with an L1 measure to calculate their difference:

$$L_{dt}{}' = \|f_{t-t}(t') - f_{s-t}(t')\|_1 \tag{5}$$

finally, the outputted loss, $L_{dt}{}'$ would be added to the final loss by the proportionally updated distillation weight, w:

$$L_{downstream-task} + w \cdot (L_{d,v} + L_{d,t} + L'_{dt}) = L_{downstream-task} + w \cdot L_d \tag{6}$$

3.4 Two-Stage Adaptive Finetuning

Domain gap exists when directly applying pretrained unimodal encoders to a specific downstream task. This may be due to the distribution differences between the pretraining and downstream data and the training task differences causing discrepancy in optimization directions. With the aim to ease the gap, we design a two-stage adaptive finetuning.

Adaptive Finetuning (AF) with Contrastive learning: We first finetune the base student VL model on a downstream dataset with Image-Text Matching (ITM). Moreover, we also incorporate the knowledge distillation on parallel as an additional task. As the base student model is adapting to the downstream dataset, the distilled knowledge from pretrained teacher models can assist the student to learn intra-sample differentiation, differences between image-text pairs. Specifically, on VCR, we convert every question and correct answer pair into a text statement with heuristic rules (details in appendix) and regard each image-text pair as the positive sample. We also apply similar procedures on VQA and SNLI-VE. For each positive sample, we create two corresponding negative samples: For creating the first negative sample, we swap the image with another randomly selected one; For the second, we swap the text.

During finetuning, we extract the representative text token feature and the full image token ([IMG]) feature to represent the text and visual sequence respectively. These two token features are then input into a prediction head, a MLP module, to predict the probability that the input image and text is a correct pair. The contrastive learning of ITM is via calculating the InfoNCE loss **?**:

$$L_{ITM} = -\log \frac{\exp\left(\text{sim}\left(z_i^I, z_i^T\right)/\tau\right)}{\exp\left(\text{sim}\left(z_i^I, z_i^T\right)/\tau\right) + \exp\left(\text{sim}\left(z_i^I, z_j^T\right)/\tau\right) + \exp\left(\text{sim}\left(z_k^I, z_i^T\right)/\tau\right)} \tag{7}$$

where $i \neq j$ and $i \neq k$. sim represents the predicted probability by the prediction head. τ is a hyperparameter temperature. z_i^I and z_i^T are the image and text representative token features for a positive pair. z_k^I and z_j^T are the image and text representative token features of other samples.

$$\begin{cases} \text{1st stage,} & L_{ITM} + w \cdot L_d \\ \text{2nd stage,} & L_{QA} + w \cdot L_d \end{cases} \tag{8}$$

Final Finetuning: Lastly, we finetune the base student model with the downstream task specific loss and also keep the knowledge distillation on parallel. The final loss is listed as above.

Method	Vision	Text	CW	TTS	AF	VCR Standard Val.		
						0.3% (100 SP/C)	3% (1000 SP/C)	100%
Baseline						30.85	53.48	75.53
Unimodal Distillation	Y					34.24	54.53	75.92
Multimodal Distillation (MD)	Y	Y				36.78	55.91	76.37
	Y	Y	Y			37.21	57.64	76.83
	Y	Y	Y	Y		38.28	58.68	77.15
Multimodal Adaptive Distillation (MAD)	Y	Y	Y	Y	Y	40.43	58.98	77.61

Table 1: VCR distillation ablation experiments using VL-BERT student model. Both vision and text teacher encoder models come from CLIP of ViT-B16. Results of low-shot experiments are averaged over more than 4 times (see supplement for standard deviations). L. represents text distillation and V. represents Vision distillation. SM = Shortcut Mitigated.

Model	MAD	Standard Evaluation				Domain-Shift Evaluation			
		0%	0.3% (100 SP/C)	3% (1000 SP/C)	100%	0%	0.3% (100 SP/C)	3% (1000 SP/C)	100%
VL-BERT	**N**	23.37	30.85	53.48	75.53	23.24	26.37	49.27	71.13
	Y	**26.80 (+3.43)**	**40.43 (+9.58)**	**58.98 (+5.5)**	**77.61(+2.08)**	**26.31 (+3.07)**	**39.27 (+12.9)**	**54.88 (+5.61)**	**74.55 (+3.42)**
UNITER	**N**	24.18	31.43	54.24	76.67	24.21	28.43	50.72	73.84
	Y	**26.78 (+2.60)**	**42.23 (+10.8)**	**59.88 (+5.64)**	**77.05 (+0.38)**	**26.49 (+2.28)**	**41.83 (+13.4)**	**54.64 (+3.92)**	**74.24 (+0.4)**
VILLA	**N**	24.22	34.84	57.01	78.27	24.62	29.41	54.15	75.43
	Y	**27.43 (+3.21)**	**43.11 (+8.27)**	**61.49 (+4.48)**	**78.91 (+0.64)**	**26.55 (+1.93)**	**41.98 (+12.57)**	**56.85 (+2.7)**	**76.32 (0.89)**

Table 2: VCR dataset results, including approaches for direct finetuning of adapters, and distillation to various student architectures. $\star$ represents our methods. Although we avoid extra pre-training, we include a recent top-performing method utilizing CLIP vision encoder with additional pre-training here for comparison (CLIP-ViL$_p$). Data sampling shown as number of image-question-answer triplets (0-shot to Full). Few-shot results are averaged over 4 runs. Refer to supplement for standard deviations, details of 0-shot experiments, and sizes of student models. Baseline refers to the original methods without distillation. Their results are based on our re-implementation of the students models with additional hyper-parameter tuning. SM = Shortcut Mitigated.

Teacher Model		VCR			SNLI-VE			VQA		
VE	TE	0.3% (100 SP/C)	3% (1000 SP/C)	100%	3000	30,000	100%	0.3% (100 SP/C)	3% (1000 SP/C)	100%
-	-	34.84	57.01	78.27	58.47	67.16	79.64	37.18	65.75	72.11
CLIP-V	CLIP-T	42.95	60.93	78.83	**59.65**	68.43	**80.67**	**40.16**	**66.93**	**73.02**
	RoBERTa	**43.11**	**61.49**	**78.91**	58.48	**68.86**	80.64	39.04	**66.23**	72.43
ViT	CLIP-T	40.73	58.38	78.59	58.08	66.12	78.37	38.79	66.10	72.02
SwinV2										
SLIP-V	SLIP-T									

Table 3: Multiple sets of teachers.

4 EXPERIMENT

4.1 MODEL

Teacher Model: For a comprehensive comparison, we experiment with a wide range of pretrained unimodal encoders. For the visual pretrained encoders, we utilize ViT, SwinV2, CLIP-V and SLIP-V; For the text pretrained encoders, we utilize RoBERTa, CLIP-T and SLIP-T.

Student Model: We adapt top-performing and publicly available VL models for highly-semantic VL tasks, including UNITER **?**, VL-BERT **?**, and VILLA **?**. All of these models represent variations of multimodal transformer-based architectures.

4.2 DATASET AND EVALUATION

For a fair comparison to demonstrate the supremacy of MAD, we evaluate MAD across the three popular highly-semantic VL datasets: VCR, VQA and SNLI-VE under three settings: zero-shot, low-shot and full-shot. Details in Appendix.

4.3 VISUAL COMMONSENSE REASONING (VCR)

Results on the VCR dataset for Q→A, across several data availability and domain, shifted constraints, are shown in Tab. 2 (for additional metrics of Q2A and Q2AR, please see supplement). These results yield 5 key observations: 1) From the 1st row, pretrained unimodal encoder models are capable of strong zero-shot results in VCR "out-of-the-box": When TE and VE are both CLIP's

vision and text encoders, the accuracy is over 58% on short-cut mitigated evaluation (a more difficult domain-shifted task), which is similar to some supervised approaches. 2) Based on the 2nd row, without enough downstream data to finetune, pretrained unimodal encoders may not adapt to the downstream domain well. In fact, new data may disturb the pretrained knowledge and result in even more poor performance than zero-shot.3) Our proposed MAD approach yields significant performance gains under low-shot data regimes across a range of student architectures: up to 52% for VL-BERT, and 47.7% for UNITER. 4) Our proposed MAD approach even benefits fully supervised tasks by 2.1% for VL-BERT and 3.8% for UNITER. 5) Our approach yields even higher gains under low-shot domain-shifted scenarios of shortcut mitigation (SM): up to 71.3% for VL-BERT and 61% for UNITER. 6) With only finetuning and not increasing any computation complexity for the student model, MAD outperforms a prior proposed approach to leverage CLIP's vision encoder and redo the pretraining steps: CLIP-Vil **?**. 7) The experiments prove that our modularized multimodal distillation method is generalizable and flexible. It can maintain effectiveness regardless of whether the pretrained vision and text encoders come from the same pretraining framework or separate ones.

MAD with VILLA delivers high performance on the public leaderboard (Q2A: **79.6%** QA2R: **82.9%** Q2AR: **66.2%**), achieving a new state-of-art of Q2A performance compared to other single models that are pretrained with image-text data, as well as overall state-of-art comparable performance for QA2R and Q2AR.

As our approach is model agnostic, ensembles are possible. Combining multiple MAD approaches, further significant gains in performance (Q2A: **80.93%**, QA2R: **84.01%**) in the fully-sampled standard validation set.

As shown in Table 1, we performed an ablation study of individual components of our MD and MAD frameworks, evaluated with VL-BERT as the student model. These results demonstrate 3 key observations: 1) Language distillation, not vision (as studied in previous works), contributes the most performance improvement. However, distilling both vision and language perform better than either one alone. 2) Multimodal distillation with an adaptive mechanism like token selection, and confidence weighting can produce better performance than naive sequence-to-sequence multimodal distillation.

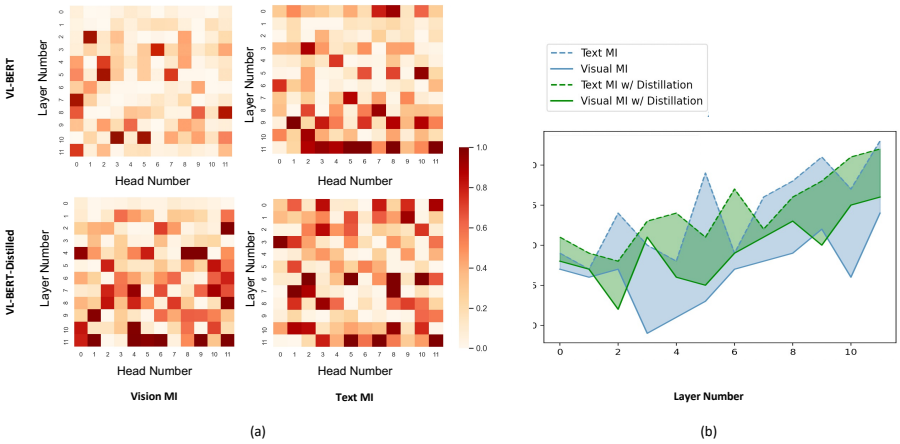

Figure 4: (a) Heatmap of MI values for both visual and textual modality. Each cell represents a combination between 12 heads and 12 layers in VL-BERT Base model. The top row corresponds to the original VL-BERT Base model and the bottom row corresponds to VL-BERT Base with Multimodal Distillation. Index starts with 0. (b) Average Modality Importance (MI) values for each layer of VL-BERT with MAD (green) and baseline (blue). Shaded areas represent differences between vision and text modalities.

Student Model	MAD	SNLI-VE					VQA				
		Val.				Test	Val.			Test-Val	Test-Std
		0 %	3000	30,000	100%	100 %	0.3% (100 SP/C)	3% (1000 SP/C)	100%		
VL-BERT	N	33.31	53.28	62.31	74.66	74.02	35.33	63.29	69.05	71.79	72.22
	Y	**35.18**	**56.78**	**65.37**	**75.75**	**75.43**	**37.12**	**65.71**	**71.42**		
UNITER	N	32.23	58.36	66.23	79.02	79.19	36.46	64.43	71.26	73.82	74.02
	Y	**33.67**	**59.42**	**68.34**	**80.14**	**80.23**	**39.75**	**65.93**	**71.94**		
VILLA	N	34.72	58.47	67.16	79.64	79.32	37.18	65.75	72.11	74.69	74.87
	Y	**36.10**	**59.65**	68.86	**80.67**	**80.32**	**40.16**	**66.93**	**73.02**	**75.81**	**76.04**

4.4 VISUAL ENTAILMENT (SNLI-VE)

Results on SNLI-VE are shown in Table **??**, revealing 3 key observations: 1) Pretrained unimodal encoders can achieve up to 40.02 % zero-shot accuracy on val set of SNLI-VE "out-of-the-box", which is significantly higher than baseline methods like VL-BERT and UNITER. 2) MAD continues to provide accuracy improvements in all settings of zero-shot, low-shot, and fully-supervised conditions, across all student base models, versus both baselines with no distillation, and MD approach. 3) MAD continues to outperform concurrent work CLIP-ViL$_p$ under all evaluated conditions.

4.5 VISUAL QUESTION ANSWERING (VQA)

Table **??** shows our results from VQA. The key observations from these experiments are as follows: 1) MAD yields performance improvement against baseline and MD under all conditions. 2) Under low-shot conditions, MAD can outperform finetuning approach CLIP-ViL$_p$ 3) Under the full-shot setting, MAD can assist the baseline method, VILLA to reach the performance of 73.02 on validation, which is comparable to the performance from CLIP-VIL$_p$. Further experimentation leveraging the model-agnostic MAD to create ensembles achieves **73.93 %** on local validation, outperforming CLIP-ViL$_p$.

4.6 ANALYSIS

Increased Utilization of Vision Modality: Following **?**, we measure Modality Importance (MI) of both modalities. After distillation, refer to Fig. 4, we observe that the Vision MI is increased in (a) and the MI difference between Vision and Text is decreased in (b).

Regulating Shortcuts: Additionally, as V+L models are prone to learn trivial shortcuts from questions to correct answers **??**, performance degrades when tested on datasets that mitigate these signals. We further seek to qualitatively better understand how our distillation approach might help improve the performance in this setting. Fig. 3, shows token attention values from a VL-BERT model on an instance from the VCR dataset, before and after distillation, in addition to the token selection scores of our MAD approach. One can see that trivial tokens such as "is" have the highest attention values prior to distillation. By contrast, the token selection forces emphasis on more meaningful terms, such as "sending," "telegram," *etc.*

5 CONCLUSIONS

In this work, we explore leveraging pretrained unimodal encoders to improve existing pretrained multimodal encoders for VL tasks. To accomplish this we present a new approach, multimodal Adaptive Distillation (MAD), to perform adaptive distillation from pretrianed unimodal encoders to a variety of high performing multimodal student encoders, including VL-BERT, UNITER, and VILLA. We evaluate this approach over a new comprehensive VL task protocol involving VCR, SNLI-VE, and VQA, covering zero-shot, low-shot, fully-supervised, and domain-shifted settings. Our results demonstrate significant improvements in performance across datasets, tasks, and settings, compared with baselines.

ACKNOWLEDGMENTS

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

REFERENCES

A APPENDIX

You may include other additional sections here.