**VIDLANKD: Improving Language Understanding via Video-Distilled Knowledge Transfer**

Zineng Tang  Jaemin Cho  Hao Tan  Mohit Bansal  
UNC Chapel Hill  
{terran, jmincho, haotan, mbansal}@cs.unc.edu

**Abstract**

Since visual perception can give rich information beyond text descriptions for world understanding, there has been increasing interest in leveraging visual grounding for language learning. Recently, vokenization [68] has attracted attention by using the predictions of a text-to-image retrieval model as labels for language model supervision. Despite its success, the method suffers from approximation error of using finite image labels and the lack of vocabulary diversity of a small image-text dataset. To overcome these limitations, we present VIDLANKD, a video-language knowledge distillation method for improving language understanding. We train a multi-modal teacher model on a video-text dataset, and then transfer its knowledge to a student language model with a text dataset. To avoid approximation error, we propose to use different knowledge distillation objectives. In addition, the use of a large-scale video-text dataset helps learn diverse and richer vocabularies. In our experiments, VIDLANKD achieves consistent improvements over text-only language models and vokenization models, on several downstream language understanding tasks including GLUE, SQuAD, and SWAG. We also demonstrate the improved world knowledge, physical reasoning, and temporal reasoning capabilities of our model by evaluating on the GLUE-diagnostics, PIQA, and TRACIE datasets. Lastly, we present comprehensive ablation studies as well as visualizations of the learned text-to-video grounding results of our teacher and student language models.

1 Introduction

Language learning can be aided by grounded visual cues, as they provide powerful signals for modeling a vastness of experiences in the world that cannot be documented by text alone [5; 29; 4]. While the recent trend of large-scale language model pretraining indirectly provides some world knowledge from text, most large text corpora (e.g., Wikipedia) do not provide enough multi-modal grounding information. Previous works have explored multiple ways of grounding language to visual information such as constructing a common vector space [38; 7] and supervising the model with token-wise generated vision labels [68]. However, the widely-used image-text datasets (e.g., MS COCO [48]) are much smaller than text-only corpora in terms of word counts and vocabulary diversity for language learning.

The recent method of ‘vokenization’ [68] is a promising initial step towards addressing this problem by supervising language models with weakly-aligned vision-language groundings. Firstly, an image-text matching model retrieves a corresponding image to each text token in a sentence. Then a language model learns to predict the selected image (called ‘voken’ ) for each text token. This can be seen as a knowledge distillation (KD) process [33] from a vision-language grounding model to a language

---

1 Code and models: https://github.com/zinengtang/VidLanKD

model. Although the voken classification task helps the language model to improve on natural language understanding (NLU) tasks, there exist several limitations: (1) images cannot faithfully convey word meanings that require more activity-based and physical commonsense knowledge. (2) the voken supervision suffers from approximation/quantization error of the text-to-image retrieval.

To address these problems, we propose a novel Video-and-Language Knowledge Distillation method, named VidLAN KD. Our teacher model consists of a video encoder and a language encoder. They are jointly trained with a video-language contrastive learning objective and a masked language modeling (MLM) objective on a multi-modal dataset (see Fig. 1). Then, we transfer the knowledge of the frozen teacher language encoder to a student language model by minimizing the distance between contextualized text representations of two models on a text dataset. For this, we propose to use different KD objectives including neuron selectivity transfer (NST) [34] and contrastive representation distillation (CRD) [71] that avoid the approximation error from voken assignments [68].

For cross-modal pretraining of our teacher model, we use HowTo100M [54], a large-scale video dataset which has more diverse vocabulary and richer world commonsense (e.g., physical and temporal) knowledge compared to MS COCO image dataset.

In our experiments, student language models learned with the proposed video-language KD objectives outperform the baseline text-pretrained language models and the models distilled with vokenization, on several diverse natural language understanding benchmarks including GLUE [73], SQuAD [61], and SWAG [79]. We also show comprehensive ablation studies on video encoders, student KD objectives, teacher pretraining objectives, and video vs. image-based pretraining. Furthermore, we empirically illustrate that our model successfully learns linguistic world knowledge and physical/temporal commonsense abilities from video, by showing improved performances on the GLUE-diagnostics [73], PIQA [6], and TRACIE [82] datasets.

Overall, our contributions are: (1) a novel cross-modal knowledge distillation method for improving natural language understanding, (2) using rich video-text data which can overcome the limitations of image vokenization, (3) empirical improvements on several language understanding benchmarks and studying different knowledge distillation methods, and (4) analysis on linguistic/physical/temporal knowledge learned from videos and ablation studies on the effectiveness of proposed components.

2 Related Work

2.1 Knowledge Distillation

Knowledge distillation (KD) [33] is the process of transferring knowledge from a teacher model to a student model. It has been successfully used in a wide range of applications, such as machine translation [40], visual recognition [31], speech recognition [10], and recommendation systems [69]. Recent works advanced the field of knowledge distillation by proposing new architectures [77; 80; 1; 55] and objectives [34; 14].

While many KD works study the problem of knowledge transfer within the same modality, cross-modal knowledge distillation [27; 20; 71] tackles the knowledge transfer across different modalities. Gupta et al. [27] transfers the knowledge of a model trained with RGB images to another model for depth maps and optical flow. Do et al. [20] proposes a KD method for visual question answering...
[2], where the trilinear (image-question-answer) relational representation of a teacher model is transferred to a bilinear (image-question) student model. Tian et al. [71] combines contrastive learning and knowledge distillation to improve the knowledge transfer between different modalities. Our ViDLANKD transfers the knowledge of a multi-modal teacher model learned from a video dataset to a student language model that tackles natural language understanding tasks.

2.2 Language Pretraining

Large-scale pretraining of contextualized language models has seen huge success in natural language processing in recent years. ELMo [57] proposes to pretrain and fine-tune a large recurrent language model, which improves performance on a diverse set of downstream natural language processing tasks. BERT [19] improves the scalability of the pretrain/fine-tune framework by using a transformer [72] language model with a masked language modeling objective. Since then, pretraining of transformer language models has been extensively explored [49; 78; 44; 22; 64; 59; 16] for various natural language understanding [61; 74; 79; 73] and generation tasks [24; 62; 61; 63].

2.3 Multi-modal Pretraining

Following the success of language pretraining with transformer models, pretraining of image-text [67; 51; 13; 47; 83; 45] and video-text [66; 54; 85; 53; 46; 70] multi-modal transformers have achieved improvements on numerous multi-modal downstream tasks [2; 76; 84]. The multi-modal transformers take both visual and textual inputs and are pretrained on image-text or video-text pairs with multi-modal masked language modeling objectives. Despite the success on multi-modal downstream tasks, Tan and Bansal [68] finds that the multi-modal pretraining does not improve (and sometimes even harms) the language understanding performance. This is because the scale and diversity of text vocabulary of image-text and video-text datasets are usually smaller than those of text datasets. To utilize the rich vocabulary of text dataset, our ViDLANKD transfers the knowledge of pretrained multi-modal model to a student language model with a large text dataset.

2.4 Visually-Grounded Language Learning

A series of works explore using visual information to aid language understanding and generation tasks including co-reference resolution [42; 15], machine translation [81], bilingual lexicon learning [39], and multi-modal contrastive learning [47]. Vokenization [68] proposes the visually-supervised language model, which is closest to our work. Vokenization proposes to supervise a language model to predict a visualized token, called ‘voken’ for each input text token. Vokens are obtained by a contextualized token-to-image matching model, pretrained on a MS COCO image captioning dataset [12]. In this work, we experiment with alternative objectives which avoid the approximation error from finite voken assignments. In addition, we use HowTo100M [54] video dataset, which provides a more diverse vocabulary as well as richer world commonsense and physical reasoning knowledge.

3 Video-Language Knowledge Distillation

3.1 Method Overview

We aim to learn a better language representation with the knowledge distilled from visual information. For this, we leverage two kinds of datasets: the aligned multi-modal dataset, $D_{VL}: \{(x, v)\}$ (e.g., HowTo100M [54]); and the text dataset, $D_L: \{x\}$ (e.g., Wikipedia), where $x$ is a sentence and $v$ is a video paired with $x$. Our knowledge transfer is done in two stages: (1) cross-modal pretraining of a teacher model, $M^T$, on multi-modal data $D_{VL}$ (Eq. 1) (2) distilling the knowledge of teacher model to a student model, $M^S$, on text data $D_L$ (Eq. 2). We illustrate our two-stage knowledge transfer method in Fig. 1.

$$\begin{align*}
\min_{\theta^T} \mathbb{E}_{x,v \sim D_{VL}} \mathcal{L}^T(M^T, x, v) \\
\min_{\theta^S} \mathbb{E}_{x \sim D_L} \mathcal{L}^{KD}(M^T, M^S, x)
\end{align*}$$

Our teacher model $M^T$ consists a language model $LM^T$ and a visual encoder $V^T$. Both $LM^T$ and $V^T$ have transformer [72] architectures, where $LM^T$ takes text tokens $x$ and $V^T$ takes video frame
3.2 Teacher Model

We train our teacher model on a multi-modal dataset with two objectives, i.e., video-language contrastive learning (Fig. 2(a)) and masked language modeling (Fig. 2(b)): $L^T = L_{CT} + L_{MLM}^2$.

Architecture. As shown in Figure 2, our teacher model $M^T$ consists of a language encoder $LM^T$ and a visual encoder $V^T$. Both $LM^T$ and $V^T$ have similar transformer architecture. For each sentence $x$, we tokenize it and append a special token [CLS] that represents the entire sentence following Devlin et al. [19]. $LM^T$ takes $x$ and outputs contextualized representation $h^x = \{h_{[CLS]}^x, h_1^x, ..., h_{|x|}^x\}$. For each video $v$, we extract frame-level features $e^v$ with an off-the-shelf image encoder (see more details in Sec. 4.2). Note that the parameters of the image encoder are not updated to save computation. We feed the frame features $e^v = \{e_1^v, ..., e_{|v|}^v\}$ to our visual encoder $V^T$ to get contextualized video frame features $h^v = \{h_1^v, ..., h_{|v|}^v\}$. We get the final video representation $h^v$ by temporally averaging frame-level features: $\overline{h^v} = \frac{1}{|v|} \sum_{i=1}^{|v|} h_i^v$. Different from Tan and Bansal [68], both $LM^T$ and $V^T$ parameters are trained from scratch.

Video-Language Contrastive Learning. To learn multi-modal grounding, we use a contrastive learning objective that encourages a closer distance between representations of aligned video-text pairs than unaligned pairs, as shown in Fig. 2(a). For each $x$, we randomly sample another text $x'$ from its batch with $x' \neq x$. Similarly, for each $v$, we randomly sample another video $v'$ from its batch.
Figure 3: Illustration of our knowledge distillation from teacher language model \( \text{LM}_T \) to student language model \( \text{LM}_S \) on a text dataset (Sec. 3.3). We train our student model with (a) knowledge distillation objectives and (b) masked language modeling.

\[
\mathcal{L}_{CT}(x, x', v, v') = \sum_i \left[ \max\{0, \alpha - \cos(h^v_i, \bar{h}^v) + \cos(h^{x'}_i, \bar{h}^{x'}) \} \right. \\
\left. + \max\{0, \alpha - \cos(h^v_i, \bar{h}^v) + \cos(h^{x'}_i, \bar{h}^{x'}) \} \right]
\]

where \( \alpha \) is the margin between the similarities of a positive pair and a negative pair. Different from previous methods [7; 32] that exploit sentence-level contrastive loss, we follow [68] to construct a token-level contrastive loss (triplet margin loss) that grounds the visual information to each contextualized token output. This fine-grained contrastive loss will help the token-level knowledge distillation in Sec. 3.4.

Masked Language Modeling. For better language understanding in our teacher model, we follow BERT [19] to use masked language modeling (MLM) objective (Fig. 2(b)). By replacing 15% of tokens in \( x \) with a special token \([\text{MASK}]\), we obtain a masked text \( x^{\text{masked}} \) with the same length. The model takes \( x^{\text{masked}} \) as input and learns to predict the tokens by minimizing the negative log-likelihoods:

\[
\mathcal{L}_{\text{MLM}}(x, x^{\text{masked}}) = -\sum_{i \in \text{Mask}} \log p(x_i | x^{\text{masked}}),
\]

where \( \text{Mask} \) refers to the indices of masked tokens.

3.3 Student Model

After we train a teacher model on a multi-modal dataset, we transfer its knowledge to a student model on a text dataset. Following Kim and Rush [40], we train our student model with a sum of masked language modeling and two knowledge distillation objectives, NST and CRD (see Sec. 3.4):

\[
\mathcal{L}^S = \mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{KD}}^{\text{NST}} + \mathcal{L}_{\text{KD}}^{\text{CRD}}
\]

Architecture. As shown in Fig. 3 (a), our student model \( M^S \) is a language model \( \text{LM}^S \) with the same transformer architecture as the teacher language model \( \text{LM}^T \). We train \( \text{LM}^S \) from scratch. Following previous works [11; 26], we introduce a multi-layer perceptron (MLP) distillation head on top of the last hidden states of \( \text{LM}^S \). In our ablation study in appendix, we find that adding a distillation head slightly improves the distillation performance.

3.4 Knowledge Distillation Objectives

We next describe the knowledge distillation (KD) objectives used in transferring knowledge from the teacher model \( \text{LM}^T \) (Sec. 3.2) to this student model \( \text{LM}^S \). Note that the weights of the teacher model \( \text{LM}^T \) are frozen during the knowledge distillation process since the teacher model should not be affected by the student model’s performance. Following Kim and Rush [40], we use the knowledge
We study the following KD objectives: Soft-label [33], L2 Regression [3], Neuron Selectivity Transfer (NST) [34], Contrastive Representation Distillation (CRD) [71], and Vokenization [68]. In our experiments comparing different KD objectives (Table 5), NST and CRD perform best, while the combination of them improved the performance even further. Therefore, we propose NST+CRD for our cross-modal knowledge distillation objective.

**Soft Label:** Hinton et al. [33] proposed a knowledge transfer method by taking a teacher model prediction with temperature scaling as a ‘soft label’. We minimize cross-entropy between $P_T(y|x)$ and $P_S(y|x)$, i.e., the word output probabilities of $LM_T$ and $LM_S$ given the input text $x$ respectively:

$$L_{\text{soft-label}}(x) = - \sum_{i=1}^{|x|} \sum_{z \in Z} P_T(y_i = z|x) \log P_S(y_i = z|x)$$

(5)

where $Z$ is the word vocabulary. Following Hinton et al. [33], we divide the softmax logits of $LM_T$ and $LM_S$ by a temperature parameter $\tau = 2.0$. Note that for soft-label KD, we reuse the LM head, instead of learning an additional distillation head.

**L2 Regression:** Following Ba and Caruana [3] which uses feature regression for KD, we minimize the squared L2 distance between $s(x)$ and $t(x)$, the last hidden states of $LM_T$ and $LM_S$ given input text $x$:

$$L_{\text{Regression}}(x) = \sum_{i=1}^{d} \|s(x)_i - t(x)_i\|^2$$

(6)

**Neuron Selectivity Transfer (NST):** NST [34] is a KD method that transfers heatmap like spatial activation patterns of teacher neurons to student neurons. We transfer the sequential activation patterns of $t(x) \in \mathbb{R}^{|x| \times d}$ to $s(x) \in \mathbb{R}^{|x| \times d}$, where $t(x)$ and $s(x)$ are the last hidden states of $LM_T$ and $LM_S$ given input text $x$, and $d$ is the hidden state dimension (# neurons). Following Huang and Wang [34], we use the squared maximum mean discrepancy (MMD) [25] with kernel trick to measure the distance between the activation patterns of student neurons $\{s(x)_{*,i}\}_{i=1}^{d}$ and teacher neurons $\{t(x)_{*,j}\}_{j=1}^{d}$:

$$\text{MMD}^2(x) = \frac{1}{d^2} \sum_{i=1}^{d} \sum_{i' = 1}^{d} k[s(x)_{*,i}; s(x)_{*,i'}] + \frac{1}{d^2} \sum_{j=1}^{d} \sum_{j' = 1}^{d} k[t(x)_{*,j}; t(x)_{*,j'}]$$

$$- \frac{2}{d^2} \sum_{i=1}^{d} \sum_{j=1}^{d} k[s(x)_{*,i}; t(x)_{*,j}]$$

(7)

where we use Gaussian kernel $k[s; t] = \exp\left(-\frac{\|s-t\|^2}{2\sigma^2}\right)$ with $\sigma = 1$. We transfer the teacher activation patterns to the student by minimizing squared MMD: $L_{\text{NST}}^\text{KD}(x) = \text{MMD}^2(x)$

**Contrastive Representation Distillation (CRD):** CRD [71] is a KD objective which maximizes the mutual information between the teacher and student representations with contrastive learning. Let’s denote $s \in S$ and $t \in T$ as student and teacher representations given $x$. We are given $N$ positive pairs (drawn from the joint distribution) for every $N$ (batch size) negative pairs (drawn from the product of marginals; independent randomly drawn inputs from $T$ and $S$). Following [71], we maximize the lower bound of mutual information between $s$ and $t$ by minimizing the following term:

$$L_{\text{CRD}}^\text{KD}(x) = -\mathbb{E}_{q(s, t|\text{positive})}[\log h(s, t)] - N \cdot \mathbb{E}_{q(s, t|\text{negative})}[\log(1 - h(s, t))]$$

(8)

where $M$ is the cardinality of the dataset, $f_1, f_2$ are learned linear layers followed by L2 normalization, which map the student and teacher representations into a same feature space. Since a large $N$ leads
to a tight mutual information lower bound, following [71], we implement a memory buffer that stores the latent features of each data sample computed from previous batches. Therefore, during training we can efficiently retrieve a large number of negative samples from the memory buffer. Note that since CRD is based on contrastive learning, it is the only KD objective where student and teacher language models can take different inputs.

**Vokenization:** Vokenization [68] could be viewed as a knowledge distillation method, where token-level text-to-image retrieval results (called ‘vokens’) of a multi-modal matching model are used as labels for a student language model. For the i-th input token $x_i$, we calculate cosine similarity between the i-th teacher language model hidden state $t(x)_i$ and a video feature $v$. Out of 30K pre-selected videos, we select a video that maximizes cosine similarity and use it as the voken for $x_i$. By denoting the voken of $x_i$ as $\text{voken}_i$, we formulate our vokenization-based KD objective as:

$$L_{\text{KD Voken}}(x) = -\sum_{i=1}^{|x|} \log P_{\text{voken}}^{S}(y_i = \text{voken}_i|x)$$  \hspace{1cm} (9)

where $P_{\text{voken}}^{S}(y|x)$ is the voken classification probabilities of LM$^S$ given input text $x$. We experiment with vokenization-based KD by retrieving vokens from images and videos (see Table 6 of Sec. 5.2). Note that vokenization suffers from approximation error; it’s hard to cover diverse textual concepts with 30K vokens. This motivates us to experiment with different ‘soft’ KD objectives described in this section (see Table 5 of Sec. 5.2).

## 4 Experimental Setup

### 4.1 Datasets

**Video-Text Dataset.** We use HowTo100M [54] for cross-modal pretraining of our teacher model (Sec. 3.2). HowTo100M has 1.22M videos totaling 136M video clips with total duration of 134,472 hours describing over 23K different visual tasks. There are 138M captions, 568M tokens with 633K distinct tokens.

**Text Pretraining Dataset.** To transfer the knowledge from our teacher language models to student language models (Sec. 3.3), we follow Tan and Bansal [68] to use English Wikipedia. For ablation studies (Sec. 5.2), we use Wiki103 [52], a widely used subset of English Wikipedia. There are 2.9B tokens and 120M sentences in English Wikipedia, and 111M tokens and 4.2M sentences in Wiki103.

**Text Downstream Dataset.** Following Tan and Bansal [68], we finetune our models on GLUE [73], SQuAD [61] 1.0 and SQuAD2.0 [60], and SWAG [79] to assess the pretraining performance. Since some smaller tasks in GLUE are reported as unstable in recent papers [21], we evaluate on the four largest datasets of GLUE: SST-2 [9], QNLI [61], QQP [35], and MNLI [74]. In addition, we also evaluate our models on the GLUE diagnostics [73], PIQA [6], and TRACIE [82] to measure its linguistic knowledge, physical reasoning, and temporal reasoning abilities.

### 4.2 Video Feature Representations

Following Miech et al. [54], we encode video features by concatenating features from a 2D frame-level image encoder and a 3D video encoder in channel dimension. Note that the parameters for 2D image encoder and 3D video encoder are not updated.

For the 2D image encoder, we sample video frames by 1fps (frame/second). The 2D image encoder outputs features for each frame individually. We experiment with ResNet-152 [30] pretrained on ImageNet-1K [18] and CLIP [58] image encoder (ViT-B/32 [23]). In contrast to conventional image encoders trained with image label classification, the CLIP image encoder is trained to match a corresponding natural language description by large-scale contrastive learning. We discuss if this natural language supervision can help our cross-modal KD in Sec. 5.1.

For the 3D video encoder, we use 3D-ResNeXt-152$^4$ [75; 28; 36] trained from a combination of publicly available datasets: ActivityNet [8], Kinetics [37], UCF-101 [65], and HMDB-51 [43]. The 3D video encoder processes 24fps videos with 3D convolution and yields features at 1.5fps. Then we sub-sample the features to 1fps to match the frame rate of 2D image encoder.

---

$^4$https://github.com/kenshohara/3D-ResNets-PyTorch
Table 1: Cross-modal knowledge distillation results of BERT\textsubscript{12L/768H} student language model on 7 downstream NLU tasks. In the first block, we include the image-based vokenization (Img-Voken) and its text-only pretrained baseline performance from Tan and Bansal [68]. In the second block, we compare our cross-modal KD method (NST+CRD) to video-based vokenization (Vid-Voken) and a text-only pretrained baseline. ‘EM’ refers to ‘Exact Match’.

<table>
<thead>
<tr>
<th></th>
<th>SST-2 Acc</th>
<th>QNLI Acc</th>
<th>QQP Acc</th>
<th>MNLI Acc</th>
<th>SQuAD v1.1 Acc</th>
<th>SQuAD v2.0 EM</th>
<th>SWAG Avg</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT\textsubscript{12L/768H} [68]</td>
<td>89.3</td>
<td>87.9</td>
<td>83.2</td>
<td>79.4</td>
<td>77.0</td>
<td>67.7</td>
<td>65.7</td>
<td>78.6</td>
</tr>
<tr>
<td>+ KD (Img-Voken) [68]</td>
<td>92.2</td>
<td>88.6</td>
<td>88.6</td>
<td>82.6</td>
<td>78.8</td>
<td>68.1</td>
<td>70.6</td>
<td>81.4</td>
</tr>
<tr>
<td>BERT\textsubscript{12L/768H}</td>
<td>89.0</td>
<td>88.0</td>
<td>86.2</td>
<td>79.2</td>
<td>77.2</td>
<td>68.0</td>
<td>65.0</td>
<td>78.9</td>
</tr>
<tr>
<td>+ KD (Vid-Voken) w/ ResNet</td>
<td>93.4</td>
<td>89.2</td>
<td>88.7</td>
<td>83.0</td>
<td>78.9</td>
<td>68.7</td>
<td>70.0</td>
<td>81.7</td>
</tr>
<tr>
<td>+ KD (Vid-Voken) w/ CLIP</td>
<td>94.1</td>
<td>89.8</td>
<td>89.0</td>
<td>83.9</td>
<td>79.2</td>
<td>68.6</td>
<td>71.6</td>
<td>82.3</td>
</tr>
<tr>
<td>+ KD (NST+CRD) w/ ResNet</td>
<td>94.2</td>
<td>89.3</td>
<td>89.7</td>
<td>84.0</td>
<td>79.0</td>
<td>68.9</td>
<td>71.8</td>
<td>82.4</td>
</tr>
<tr>
<td>+ KD (NST+CRD) w/ CLIP</td>
<td>94.5</td>
<td>89.6</td>
<td>89.8</td>
<td>84.2</td>
<td>79.6</td>
<td>68.7</td>
<td>72.0</td>
<td>82.6</td>
</tr>
</tbody>
</table>

4.3 Implementation Details

For the student distillation head, we use a two-layer MLP with ReLU activation. For both student and teacher language models, following previous works [49; 17; 68], we truncate input text that is longer than 128 tokens. We truncate videos features that are longer than 512 frames. We use an AdamW [41] optimizer with learning rate 2e-4 and weight decay [50] of 0.01. We reserve 10K samples of the HowTo100M dataset as validation data. We train the teacher model until it converges on validation data. For downstream tasks, we report the results on the validation sets. We train 3 epochs with a learning rate of 1e-4 and a batch-size of 32 for all downstream tasks. We use hinge loss margin $\alpha = 1.0$ for $\mathcal{L}_{CT}$ (Eq. 3). We implement our models with PyTorch 1.5 [56] and train them with Nvidia GeForce RTX 2080ti GPUs. For teacher pretraining, we use 4 GPUs for BERT\textsubscript{12L/768H} and BERT\textsubscript{6L/512H} models for 7 days and 2.5 days respectively. For knowledge distillation, we use 4 GPUs for BERT\textsubscript{12L/768H} and BERT\textsubscript{6L/512H} models for 10 days and 3 days respectively.

5 Results and Analysis

5.1 Primary Downstream Task Results

In the first block of Table 1, we include the image-based vokenization (Img-Voken) and their text-only pretrained baseline from Tan and Bansal [68]. Given our reproduced text-only baseline shows a similar average performance (78.9 vs 78.6), our student models distilled from NST+CRD are much better (82.6 vs 81.4). We discuss the comparison between video-based and image-based KD in detail in the following ablation study in comparison to vokenization (Table 6).

In the second block of Table 1, we compare our proposed cross-modal KD method (NST+CRD) to video-based vokenization (Vid-Voken) and a non-KD baseline (BERT\textsubscript{12L/768H}) which is only pretrained on text. We can see both cross-modal KD methods (i.e., KD and Vid-Voken) significantly outperform the text-only baseline across all 7 downstream tasks. We also experiment with different 2D frame encoders (Sec. 4.2): ResNet and CLIP. For both Vid-Voken and NST+CRD, we observe CLIP further improves the performance results over ResNet, indicating using a strong visual encoder helps the teacher training and thus benefits the knowledge distillation.

5.2 Ablation Studies

In this section, we conduct a comprehensive ablation study to show the effectiveness of our proposed methods. For all ablation experiments, we use BERT\textsubscript{6L/512H} architecture for student and teacher language models. We use ResNet-152 for 2D frame encoder and 3D-ResNeXt-152 for 3D frame encoder (Sec. 4.2). Wiki103 [52] is used for student model training. We also perform ablation experiments on the effect of additional distillation head in appendix.

---

\textsuperscript{3}Vokenization uses pretrained BERT checkpoint for its ‘teacher’ (vokenizer) model but we train our teacher language model fully from scratch.
Text-only Pretraining. Our cross-modal KD improves the performance on downstream NLU tasks significantly (Sec. 5.1). Where does the improvement come from, video or text? To answer this question, we conduct text-only pretraining of BERT$_{6L/512H}$ on Wiki103 text (111M tokens), HowTo100M captions (568M tokens) and compare them to a no-pretrain baseline. In Table 2, while both pretrained models improve the performance over the no-pretrain baseline, Wiki103-trained model outperforms HowTo100M-trained model (which has more tokens) significantly. This indicates that our KD methods improve NLU performance because of multimodal grounding, instead of just the larger corpus.

Effect of Teacher Training Objectives. We here analyze the teacher training objectives by comparing the corresponding distilled student model results. In Table 3, the teacher model trained solely with MLM (+KD from $T^{MLM}$) does not significantly change the student model performance. At the same time, the teacher model trained with only visual supervision, i.e., contrastive objective (+KD from $T^{CT}$), improves the result. This illustrates the motivation to perform knowledge transfer from a visually supervised MLM model. Lastly, combining the MLM and the contrastive objective (+KD from $T^{MLM+CT}$) in teacher model training shows the best student results.

Two-stage PT vs. Cross-modal KD. In Table 4, we compare two-stage pretraining with a single model to our proposed cross-modal KD approach. For single model baselines, we use text-only (MLM on Wiki103), video-only (MLM+CT on HowTo100M), and two-stage (video-then-text) pretraining. While the two-stage pretraining shows better results than the text/video-only pretraining, our VidLanKD outperforms all baselines on GLUE tasks, especially on SST-2 and MNLI.

KD Objectives Comparison. In Table 5, we compare different knowledge distillation objectives introduced in Sec. 3.4. The student models trained with NST [34] and CRD [71] show the best finetuning performance on downstream tasks. When combining NST and CRD, performance further improves with marginal additional computation cost, hence we propose to use NST+CRD for our cross-modal knowledge distillation.

Comparison to Vokenization. In Table 6, we compare NST [34] and Vokenization [68] in both image and video-level teacher model supervision. For video-level supervision, we provide our visual encoder with the whole video features (Sec. 4.2). For image-level supervision, we provide our visual encoder only with 2D features of the middle frame for each video clip. With image-level super-

---

Table 2: Text-only pretraining results of BERT$_{6L/512H}$ pretrained on Wiki103, HowTo100M captions, and no-pretrain baseline.

<table>
<thead>
<tr>
<th>Pretrained on</th>
<th>SST-2</th>
<th>QNLI</th>
<th>QQP</th>
<th>MNLI</th>
</tr>
</thead>
<tbody>
<tr>
<td>No-Pretrain</td>
<td>79.6</td>
<td>61.5</td>
<td>72.7</td>
<td>61.6</td>
</tr>
<tr>
<td>Wiki103 (Formal language)</td>
<td><strong>88.8</strong></td>
<td><strong>84.9</strong></td>
<td><strong>85.3</strong></td>
<td><strong>77.4</strong></td>
</tr>
<tr>
<td>HowTo100M (ASR captions)</td>
<td>83.3</td>
<td>78.5</td>
<td>83.7</td>
<td>71.5</td>
</tr>
</tbody>
</table>

Table 3: Ablation results showing the effect of the teacher model’s training objectives. NST is used for knowledge distillation.

<table>
<thead>
<tr>
<th>Model</th>
<th>SST-2</th>
<th>QNLI</th>
<th>QQP</th>
<th>MNLI</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT$_{6L/512H}$</td>
<td>88.8</td>
<td>84.9</td>
<td>85.3</td>
<td>77.4</td>
</tr>
<tr>
<td>+KD from $T^{MLM}$</td>
<td>88.1</td>
<td>83.1</td>
<td>85.6</td>
<td>77.4</td>
</tr>
<tr>
<td>+KD from $T^{CT}$</td>
<td>88.9</td>
<td><strong>85.2</strong></td>
<td>86.2</td>
<td>77.5</td>
</tr>
<tr>
<td>+KD from $T^{MLM}+CT$</td>
<td><strong>91.1</strong></td>
<td>85.0</td>
<td><strong>87.4</strong></td>
<td><strong>78.4</strong></td>
</tr>
</tbody>
</table>

Table 4: Comparison of pretraining on text, video, both (Two-stage PT), and our VidLanKD.

<table>
<thead>
<tr>
<th>Model</th>
<th>SST-2</th>
<th>QNLI</th>
<th>QQP</th>
<th>MNLI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text PT</td>
<td>88.8</td>
<td>84.9</td>
<td>85.3</td>
<td>77.4</td>
</tr>
<tr>
<td>Video PT</td>
<td>84.0</td>
<td>78.9</td>
<td>84.2</td>
<td>73.1</td>
</tr>
<tr>
<td>Two-Stage PT</td>
<td>90.3</td>
<td><strong>85.0</strong></td>
<td>87.2</td>
<td>76.9</td>
</tr>
<tr>
<td>VidLanKD</td>
<td><strong>91.1</strong></td>
<td><strong>85.0</strong></td>
<td><strong>87.4</strong></td>
<td><strong>78.4</strong></td>
</tr>
</tbody>
</table>

Table 5: Ablation of knowledge distillation objectives.

<table>
<thead>
<tr>
<th>SST-2</th>
<th>QNLI</th>
<th>QQP</th>
<th>MNLI</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT$_{6L/512H}$</td>
<td>88.8</td>
<td>84.9</td>
<td>85.3</td>
</tr>
<tr>
<td>+KD-Soft label</td>
<td>87.2</td>
<td>84.4</td>
<td>86.4</td>
</tr>
<tr>
<td>+KD-Regression</td>
<td>88.8</td>
<td>84.8</td>
<td>87.1</td>
</tr>
<tr>
<td>+KD-Vid Voken</td>
<td>89.7</td>
<td>85.5</td>
<td>86.5</td>
</tr>
<tr>
<td>+KD-NST</td>
<td><strong>91.1</strong></td>
<td>85.0</td>
<td><strong>87.4</strong></td>
</tr>
<tr>
<td>+KD-CRD</td>
<td>90.0</td>
<td>85.5</td>
<td>87.3</td>
</tr>
<tr>
<td>+KD-NST+CRD</td>
<td><strong>91.5</strong></td>
<td><strong>85.8</strong></td>
<td><strong>87.4</strong></td>
</tr>
</tbody>
</table>

Table 6: Comparison between vokenization (Voken) and NST with image and video-level supervision.

<table>
<thead>
<tr>
<th>SST-2</th>
<th>QNLI</th>
<th>QQP</th>
<th>MNLI</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT$_{6L/512H}$</td>
<td>88.8</td>
<td>84.9</td>
<td>85.3</td>
</tr>
<tr>
<td>+KD-Voken (Image)</td>
<td>89.3</td>
<td>84.4</td>
<td>86.0</td>
</tr>
<tr>
<td>+KD-NST (Image)</td>
<td>88.9</td>
<td>85.0</td>
<td>86.3</td>
</tr>
<tr>
<td>+KD-Voken (Video)</td>
<td>89.7</td>
<td><strong>85.5</strong></td>
<td>86.5</td>
</tr>
<tr>
<td>+KD-NST (Video)</td>
<td><strong>91.1</strong></td>
<td>85.0</td>
<td><strong>87.4</strong></td>
</tr>
</tbody>
</table>
vision (first block), Vokenization and NST show comparable performance. However, with video-level supervision (second block), NST outperforms Vokenization on 3 out of 4 tasks. The gap in the video domain might come from voken approximation error, where each image or video input is approximated with one of 30K predefined vokens. Since videos usually contain more diverse contents than images, the voken approximation error would be amplified in video-level supervision, whereas our NST distillation avoids this issue.

### 5.3 Analyzing the Knowledge Learned from Video

In this subsection, we analyze the knowledge that our language models learn from video via cross-modal knowledge distillation. To measure linguistic knowledge and physical/temporal reasoning ability, we show results of our models on the GLUE diagnostics [73], the Physical Interaction Question Answering (PIQA) [6], and TRACIE [82]. In addition, we visualize the learned multi-modal grounding ability of our model with text-to-video retrieval.

Table 7: Finetuning performance on GLUE diagnostics [73], PIQA [6] and TRACIE [82] datasets, which measure the linguistic knowledge, physical and temporal reasoning capabilities of language models, respectively.

<table>
<thead>
<tr>
<th>GLUE diagnostics</th>
<th>PIQA</th>
<th>TRACIE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexicon</td>
<td>53.0</td>
<td>64.2</td>
</tr>
<tr>
<td>Predicate</td>
<td>64.2</td>
<td>53.3 (+0.3)</td>
</tr>
<tr>
<td>Logic</td>
<td>44.5</td>
<td>63.7 (-0.5)</td>
</tr>
<tr>
<td>Knowledge</td>
<td>44.0</td>
<td>48.6 (+4.6)</td>
</tr>
</tbody>
</table>

Linguistic Knowledge. GLUE diagnostics dataset [73] evaluates sentence understanding through natural language inference (NLI) problems. The dataset consists of sentence pairs labeled with their entailment relations (entailment, contradiction, or neutral) in both directions and tagged with a set of entailment labels. Each example in the dataset is labeled with 4 categories of linguistic phenomena: (1) lexical semantics, (2) predicate-argument structure, (3) logic, and (4) knowledge (including common sense). In Table 7, we compare the baseline language model (BERT\(_{6L/512H}\)) to our NST-distilled model. We finetune the models on MNLI [74] that has the same format and test on GLUE diagnostics. We observe a large gain on the knowledge category (which involves common sense and external world knowledge) while there are no significant differences on other categories. This suggests that our student model learns the external, grounded world knowledge in the teacher model and the video-text dataset.

Physical and Temporal Reasoning. PIQA [6] is a question answering dataset evaluating physical interactions and commonsense reasoning. TRACIE [82] is a temporal reasoning benchmark on implicit events, which are not mentioned explicitly in natural language text but can be inferred from it. In Table 7, our BERT\(_{6L/512H}\) distilled with NST significantly outperform the text-only pretrained baseline on both benchmarks. The finding suggests (consistent with the GLUE diagnostics findings above) that video knowledge distillation also helps improve the physical and temporal reasoning capabilities of the language model. See appendix for the more detailed discussion on the PIQA and TRACIE experiment.

### 6 Conclusion

We introduce V\(_{IDL}AN\)KD, a novel cross-modal knowledge distillation method to help general language understanding. Our teacher model is first trained on a video-text dataset, and then we transfer its knowledge to a student language model with a text dataset. Via the distillation objectives and video-text datasets, our method overcomes the limitations of the recent vokenization method. We empirically demonstrate that our V\(_{IDL}AN\)KD improves on several NLU tasks over models trained by pure-text or vokenization. We conduct comprehensive ablation analysis to show the effectiveness of each proposed component. We also illustrate the linguistic knowledge and physical/temporal commonsense reasoning learned from videos, and visualize our model’s multi-modal grounding ability.
Acknowledgments

We thank the reviewers for their helpful comments. We thank Yixin Nie and Gabriel Ilharco for useful dataset suggestions. This work was supported by ARO-YIP Award W911NF-18-1-0336, DARPA MCS Grant N66001-19-2-4031, DARPA KA IOS Grant FA8750-19-2-1004, Google Focused Research Award, and Bloomberg Data Science Ph.D. Fellowship. The views, opinions, and/or findings contained in this article are those of the authors and not of the funding agency.

References


[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In CVPR.


[61] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In EMNLP.


Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In ICCV.


Adina Williams, Nikita Nangia, and Samuel R Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In NAACL.

Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In ECCV.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In CVPR.

Zheng Xu, Yen-Chang Hsu, and Jiawei Huang. 2018. Training shallow and thin networks for acceleration via knowledge distillation with conditional adversarial networks. In ICLR Workshop.


