SINC: Self-Supervised In-Context Learning for Vision-Language Tasks

Yi-Syuan Chen[1], Yun-Zhu Song[1], Cheng Yu Yeo[1], Bei Liu[2], Jianlong Fu[2], and Hong-Han Shuai[1]
[1]National Yang Ming Chiao Tung University, [2]Microsoft Research Asia
{yschen.ee09,yzsong.ee07,boyyeo123.ee08,hhshuai}@nycu.edu.tw
{Bei.Liu,jianf}@microsoft.com

Abstract

*Large Pre-trained Transformers exhibit an intriguing capacity for in-context learning. Without gradient updates, these models can rapidly construct new predictors from demonstrations presented in the inputs. Recent works promote this ability in the vision-language domain by incorporating visual information into large language models that can already make in-context predictions. However, these methods could inherit issues in the language domain, such as template sensitivity and hallucination. Also, the scale of these language models raises a significant demand for computations, making learning and operating these models resource-intensive. To this end, we raise a question: "How can we enable in-context learning without relying on the intrinsic in-context ability of large language models?". To answer it, we propose a succinct and general framework, **S**elf-supervised **IN**-**C**ontext learning (SINC), that introduces a meta-model to learn on self-supervised prompts consisting of tailored demonstrations. The learned models can be transferred to downstream tasks for making in-context predictions on-the-fly. Extensive experiments show that SINC outperforms gradient-based methods in various vision-language tasks under few-shot settings. Furthermore, the designs of SINC help us investigate the benefits of in-context learning across different tasks, and the analysis further reveals the essential components for the emergence of in-context learning in the vision-language domain.*

1. Introduction

Large language models such as GPT-3 [6] are able to perform *in-context learning (ICL)*: given a prompt consisting of a series of demonstrations and a query data as input, the model can generate the corresponding prediction without any parameter updates. Meanwhile, recent works show that large vision-language (VL) models [2, 67] can also possess such an ability. Specifically, these models can rapidly incorporate multimodal information with few demonstrations for tackling a variety of downstream tasks, such as image captioning [9], visual question answering [22], and fast concept binding [6]. Behind the success, the shared scheme of these approaches is to incorporate visual information into large language models via proposed modules. In particular, the in-context ability of these VL models would significantly rely on the language side. As such, the issues in the language domain could be inherited, such as template sensitivity [40, 52] and hallucination [27]. Previous studies [19, 2, 6, 56, 67] also indicate that the in-context ability scales with the model sizes and barely emerges in smaller models [6]. This property requires current methods to be built upon large language models for leveraging in-context demonstrations. Although previous works typically freeze language models for training efficiency, the language mod-

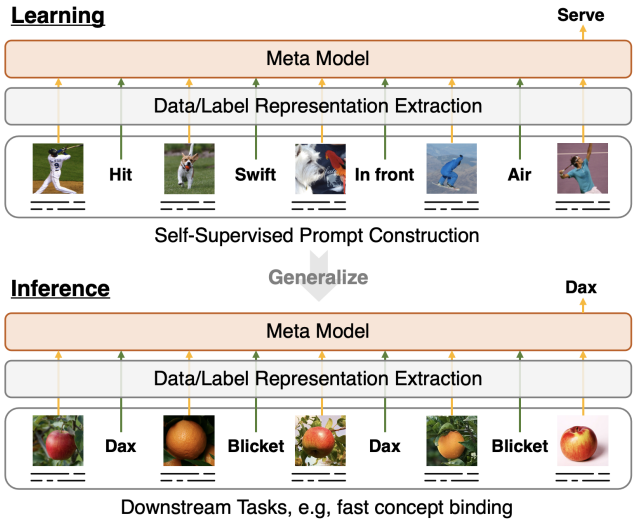

Figure 1. **Illustration of SINC.** A meta-model is introduced for acquiring in-context ability given features extracted from general models. During pre-training, the meta-model is learned on prompts constructed in a self-supervised manner. Our designs jointly enable the transfer of in-context ability to the downstream.

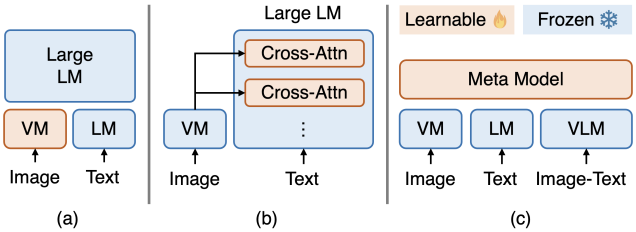

Figure 2. **Architectural comparison.** Previous works (a) [67] and (b) [2] achieve in-context learning for VL tasks with large language models. Our SINC relieves such a constraint by introducing a meta-model for acquiring the in-context ability.

els are still involved in the learning process, creating a non-negligible demand for resources to learn or operate these models [64]. Moreover, the length of demonstrations could readily exceed the practical limitation of most Transformer models [78], especially for vision-language data.

To this end, we pose a challenging research question: "*How can we enable in-context learning without relying on the intrinsic in-context ability of large language models?*" The access to the solution could lie in the understanding of ICL properties in large language models. Previous studies have shown that the formats of demonstrations can affect performance drastically [37, 40, 79, 81]. Furthermore, [43] shows that randomly assigning labels for the demonstrations could barely decrease the performance, while [77] justifies that the phenomenon is not generalized across tasks. In view of these obscure observations, recent works attempt to study ICL specifically from different perspectives. [1, 14] show that Transformers trained from scratch can implicitly implement the gradient descent algorithm in their forward pass, and the number of attention layers [82] could further relate to the equivalent learning steps. [7, 56], in another way, study the effects of training data for language models. The results reveal that the emergence of ICL could be attributed to certain language properties such as burstiness [33, 54]. Overall, these studies suggest that the in-context ability of language models results from multiple factors. Moreover, *this ability could be incidental since typical language modeling is not intentionally designed based on these factors*. Thus, models could exhibit unexpected behaviors, and the acquisition of in-context ability could be inefficient in terms of model capacity.

To address previous limitations, we present a general framework, named **S**elf-supervised **IN**-**C**ontext learning (SINC). The core idea is to decouple the acquisition of in-context ability from conventional VL pre-training and incentivize it through both architectural and data perspectives. Specifically, we introduce a *meta-model* as the in-context learner that directly operates on the representations produced from frozen models. A self-supervised learning scheme is proposed to enable the meta-model to make predictions based on demonstrations, with the transferability across tasks. In particular, we learn the meta-model on tailored prompts comprising sequences of data and label representations. For constructing data-label pairs in the prompts, inspired by the literature of question answering [20, 50, 71], we regard data sharing similar missing semantics as homogeneous and group them as a class. This strategy enables us to create diverse labels from unannotated image-text pairs. However, we identify that the predictions from models would be agnostic to the demonstrations if there is no adequate correlation with the query data. Therefore, we leverage the idea from few-shot learning [11, 62] to create specific prompts to trigger the model for utilizing the demonstrated information. Furthermore, in our observations, downstream tasks would demand the in-context ability to different degrees. We thus propose learning different prompts with a controllable ratio to better benefit and study different tasks. Regarding the formation of representations, on the data side, we propose incorporating pre-trained models from various domains, where the produced features are further aggregated with the proposed *multi-source feature fuser (MFF)*. On the label side, the representations are composed of subword embeddings [53] of label descriptions, enabling the generalization to unseen labels. Overall, our representation-level in-context learner could be transferred to different scenarios after pre-training, as shown in Fig. 1. A comparison of our architecture with prior works [2, 67] is depicted in Fig. 2, wherein prior works either (a) prepend a learnable vision encoder or (b) interleave adapter modules to the large language models for ICL. In contrast, we achieve ICL by introducing the meta-model after the frozen models, enabling us to prepare the representations on separate devices or in an offline manner. This scheme exempts the frozen models from all the backward processes, thereby significantly alleviating the computation burden. The main contributions of this paper are summarized as follows.

- We propose a novel framework, SINC, that decouples the acquisition of ICL from VL pre-training, enabling ICL in a more manageable and extensible way without relying on the intrinsic in-context ability of large language models.
- We propose to learn a meta-model on self-supervised prompts consisting of tailored demonstrations. The learned models can be transferred to downstream tasks for making in-context predictions on-the-fly.
- Extensive experiments show that SINC outperforms previous gradient-based methods and a strong ICL baseline. The analysis further reveals the properties and essential components for ICL in the VL domain.

2. Related Works

Vision-Language Pre-training. Pre-training for vision-language scenarios is a rapidly evolving domain that aims to bridge the gap between visual perception and natural language comprehension. Present methodologies predominantly employ large-scale transformer-based models, which have showcased impressive effectiveness [74, 63, 72, 73, 24, 25]. Various approaches from distinct perspectives have been proposed for enhancement, such as learning objectives [26, 75], frozen-model utilization [34, 2], visual representations [24], alignments [73, 74], and pre-training datasets [72, 35]. Moreover, a research direction has emerged to further leverage these models in scenarios with limited resources. This research line concentrates on integrating lightweight modules, such as adapters [11, 12], to enable efficient fine-tuning of vision-language models [65, 64, 80]. However, it's important to note that these techniques necessitate alterations to the model architecture, followed by subsequent fine-tuning, which may not be suitable for situations where the model remains inaccessible.

In-Context Learning. Since [6] demonstrated the emergence of the in-context learning (ICL) ability in large-scale language models (LLMs), there has been a growing interest in utilizing the ICL paradigm [23, 37, 40, 41, 44, 51, 52, 81]. A research line based on pre-trained LLMs has emerged to explain the mechanism of ICL through the lens of pre-training data [51, 56, 7], in-context examples [43, 77], and model architecture [5, 45]. Studies suggest that the behavior of in-context learners is driven by the distributions of pre-training data, such as burstiness [33, 7, 54] and numbers of rarely occurring classes [51]. Combining different training corpora can also facilitate the emergence of ICL [56]. Further research has found that the label space and input text distribution are more crucial than providing correct labels for demonstrations [43], while [77] justifies that the observations could be limited in specific tasks. Additionally, induction heads in large Transformer models may contribute to ICL [45], and only a few nucleus layers are essential across downstream tasks, suggesting that LLMs may be under-trained [5]. To incorporate ICL into a vision-language model (VLM) [74, 63, 72, 73, 24, 25], [67] proposes encoding images into the word embedding space of an LLM, while [2] achieves this by interleaving proposed modules to an LLM. However, the potential pitfalls of LLM-based ICL is that LM pre-training is not explicitly designed for this task, and the ability of ICL is therefore implicitly learned as a by-product. Thus, further warmup, calibration, or template designs are usually required [8, 16, 23, 37, 52, 81]. Specifically, to bridge the gap between LM pre-training and downstream tasks, [10, 42] apply meta-training and [30] explores the prompt-based tuning. These studies indicate that LLMs are not the sole approach for obtaining ICL ability, and a thorough comprehension of underlying factors is critical. Consequently, aside from the LLM-based ICL mentioned above, some works focus on investigating the empirical properties of ICL, such as task construction or model architecture. For instance, [19] shows that Transformer models trained from scratch can in-context learn the class of linear functions. [69] finds that training Transformers on auto-regressive tasks is closely related to gradient-based meta-learning formulations. [1] shows that Transformer-based in-context learners implement standard learning algorithms implicitly for linear regressions. Overall, these studies investigate the ICL ability from different perspectives of LLMs to enhance understanding, but most are limited to crafted datasets or simplified architectures. Thus, building on these efforts, we aspire to expand ICL research to more realistic scenarios.

Multimodal Few-shot Learning. Few-shot learning has gained prominence in recent years with the rise of pre-trained language models [59]. To extend this capability to a multimodal setting, some prior works incorporate lightweight modules, such as adapters [11, 12], to enable efficient fine-tuning [65, 64, 80] with limited data. On the other hand, [67] prepends a trainable vision encoder to a frozen GPT-like LM with 7 billion parameters for ICL. Similarly, [2] interleaves trainable adapter modules to a frozen LM of 70B parameters and uses in-context examples as prompts. To leverage external knowledge for few-shot learning with GPT-3 [6], [76] converts images to captions to utilize textual demonstrations for ICL. However, the use of large pre-trained VL models could be impractical for real-world applications due to their size. To this end, [28] examines the effect of prompts and pre-training objectives on relatively smaller few-shot learners. Notably, previous works mainly focus on question-answering or generative tasks, neglecting other reasoning tasks [61, 70].[1] These tasks are challenging due to their complex semantics, making adaptation from few examples difficult. Thus, we further include these tasks to extensively evaluate our methods.

3. Methodology

In this section, we introduce the self-supervised in-context learning (SINC). We first describe the formulation of our methods, and then introduce the overall framework Finally, we delineate the crucial details of developing the in-context ability in a self-supervised manner based on our framework. The overview of SINC is shown in Fig. 3.

3.1. Formulation

Let $\mathcal{F}$ be a set of vision-language tasks and $f \in \mathcal{F}$: $\mathcal{X}_f \rightarrow \mathcal{C}_f$ is a mapping function, where $\mathcal{X}_f$ is the set of input data and $\mathcal{C}_f$ is the set of classes. A prompt π on task

[1] [58] proposes a zero-shot approach for SNLI-VE. However, the method is closely tied to additional annotations, as detailed in Appendix.

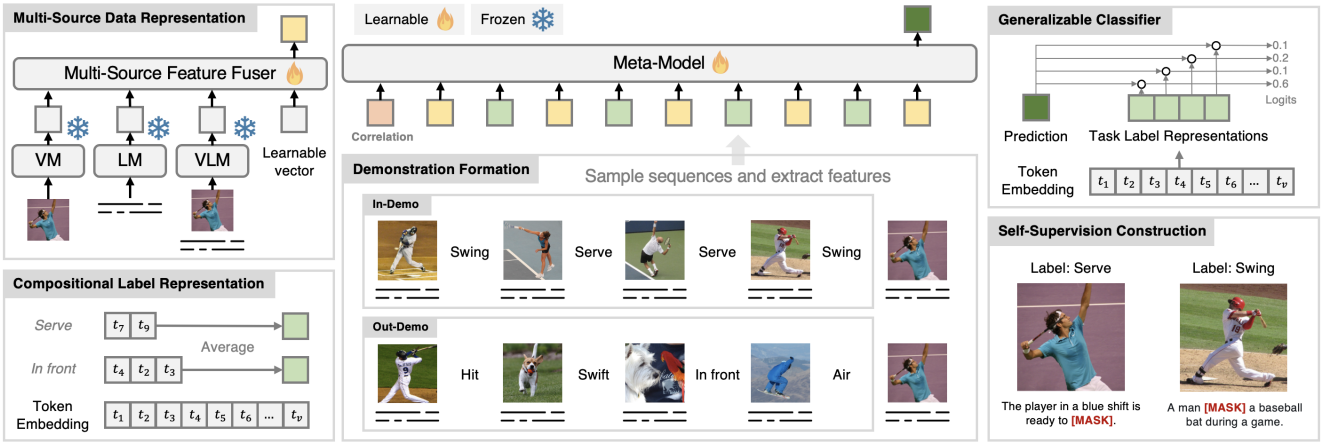

Figure 3. **Framework of SINC.** The meta-model learns on prompts comprising a sequence of data and label representations. Data representation is extracted from multiple pre-trained models (top-left). Label representation is assembled compositionally through pre-trained token embeddings (bottom-left). To incentivize the in-context ability, we construct prompts with tailored demonstrations (bottom-center) in a self-supervised manner (bottom-right). Prediction is then conducted by a classifier that could generalize across tasks (top-right).

f is a sequence $(x_1^d, f(x_1^d), ..., x_n^d, f(x_n^d), x^q)$ consisting of a series of demonstrations $\{x_i^d\} \subset \mathcal{X}_f$ and a query data $x^q \in \mathcal{X}_f$. Consider a model θ, we say the model can in-context learn up to ϵ if it can predict $f(x^q)$ as:

$$\mathbb{E}_{\mathcal{F},\mathcal{X}_f}[\mathcal{L}(\theta(\pi), f(x^q)) - \mathcal{L}(\theta(x^q), f(x^q))] \leq \epsilon, \quad (1)$$

where $\mathcal{L}$ is an appropriate loss function depending on f. Specifically, Eq. 1 indicates whether predictions of a model can be improved with demonstrations. We aim to learn such a model θ with a pretext task f^{src} and generalize to downstream tasks $\{f_i^{tgt}\} \subset \mathcal{F}$. Furthermore, to investigate the properties of models and tasks, we define the *in-context benefit (ICB)* for a task f on a model θ as:

$$ICB(\theta, f) \doteq \mathbb{E}_{\mathcal{X}_f}[\frac{\mathcal{L}(\theta(x^q), f(x^q)) - \mathcal{L}(\theta(\pi), f(x^q))}{\mathcal{L}(\theta(x^q), f(x^q))}], \quad (2)$$

which evaluates the ratio of performance improvements for a model utilizing in-context demonstrations.

3.2. Overall Framework

Architecture. SINC learns a model θ comprising a *base model* θ^{base} and a *meta-model* θ^{meta}. Given a prompt $\pi = (x_1^d, f(x_1^d), ..., x^q)$, we utilize the base model to extract the representations for the sequence, denoted as $(h_1^d, \widehat{h}_1^d, ..., h^q)$. Now, given the representations of preceding demonstrations, we learn the model to predict $f(x^q)$ by minimizing the expected loss:

$$\begin{aligned}\mathcal{L} &= -\mathbb{E}_{\mathcal{X}_f}[\log P(f(x^q)|\pi, \theta^{base}, \theta^{meta})] \\ &= -\mathbb{E}_{\mathcal{X}_f}[\log P(f(x^q)|h^q, \{h_i^d\}, \{\widehat{h}_i^d\}, \theta^{meta})].\end{aligned} \quad (3)$$

The meta-model is a decoder-only Transformer [49, 6], and we only consider the prediction of $f(x^q)$ for loss computation. Next, we address the constructions of h and $\widehat{h}$.

Multi-Source Data Representations. We design our base model θ^{base} to flexibly cooperate with multiple knowledge sources. Specifically, given the vision-language data x and several pre-trained models $\{\phi_i\}$ specialized in different modalities, e.g., vision models, language models, and vision-language models, we first extract data features from each of the models with respect to the corresponding modalities, denoted by $\{z_i\}$. Next, we propose a *multi-source feature fuser (MFF)* ϕ^{mff} and a learnable indicator z' to aggregate the multimodal information from different sources. The data representation h is then obtained as follows:

$$h = h_{z'} = \phi^{mff}(z', \{z_i\}), \quad (4)$$

where ϕ^{mff} is composed of cross-attention layers [68] and $h_{z'}$ is the output hidden state of z'. Overall, the base model θ^{base} comprises the pre-trained models $\{\phi_i\}$ and the multi-source feature fuser ϕ^{mff}. Importantly, we keep $\{\phi_i\}$ frozen throughout the learning, which significantly reduce the computation demands. Moreover, attributed to the design of our architecture, $\{z_i\}$ can be prepared offline, as the pre-trained models only need to be forwarded once and can be exempted from all the backward processes, thus achieving better efficiency than the previous method [67].

Compositional Label Representations. One viable strategy to construct label representations is to create a learnable embedding for each label from scratch. However, such an approach requires maintaining specific embeddings for different tasks, which could impede the model generalizability in downstream tasks since some labels could be unseen during pre-training. To this end, we propose creating label representations in a *compositional* way. The core idea is to leverage pre-trained token embeddings from the base model θ^{base}. Specifically, consider such a token embedding $E \in \mathbb{R}^{|\mathcal{V}| \times m}$, where $\mathcal{V}$ is the vocabulary and m is the em-

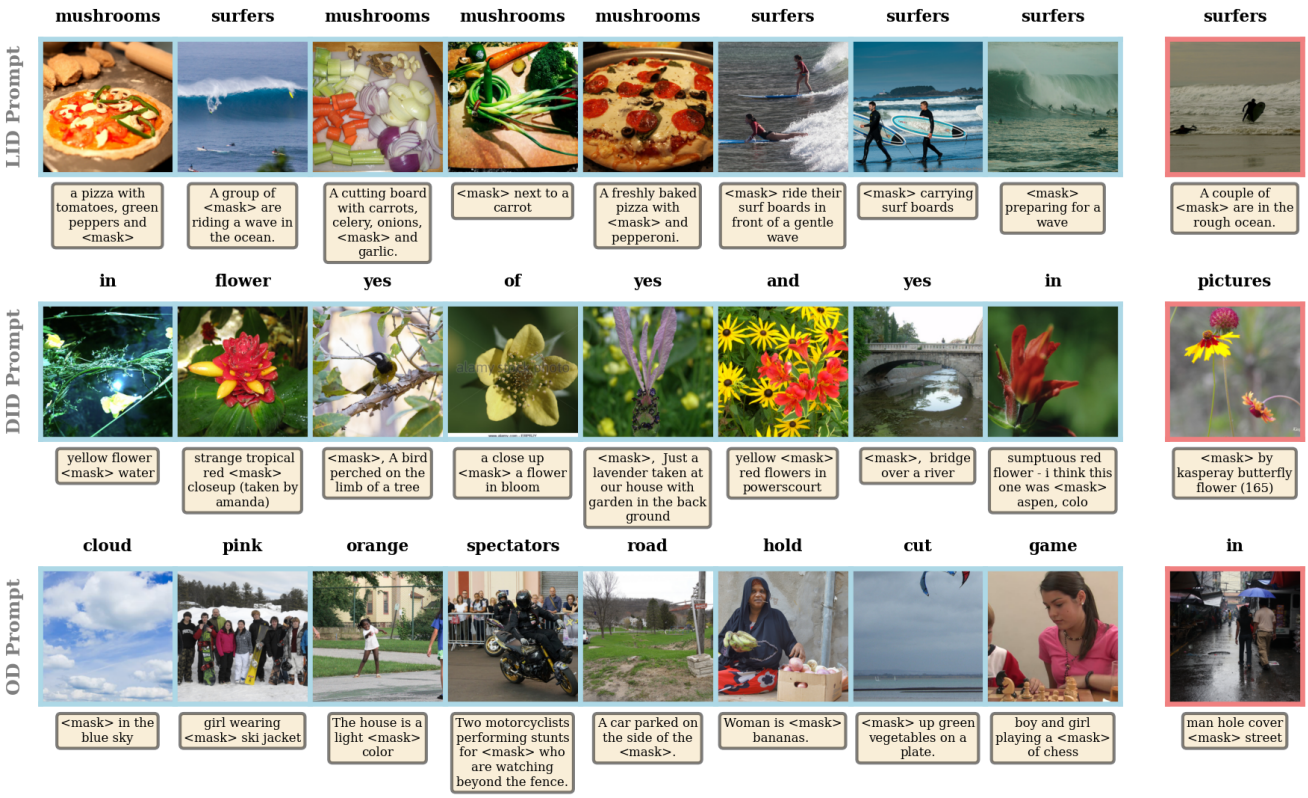

Figure 4. Examples of Label-In-Demo (LID), Data-In-Demo (DID) and Out-Demo (OD) prompts in pre-training. Demonstrations are outlined in blue and query data is outlined in red. Images are center-cropped for better visualization.

bedding size. We tokenize the descriptions of a label $f(x)$ into a sequence $(t_1, ..., t_k)$. The label representation is then constructed by averaging over the token embeddings of the sequence as $\widehat{h} = \frac{1}{k}\sum_i E_{t_i}$, where $E_{t_i} \in \mathbb{R}^m$ is the embedding of token t_i. This design allows the model to generalize to various unseen labels in downstream.

Generalizable Classifier. Following the idea of label representation construction, the classifier should also be generalizable across different tasks. In this regard, we propose to utilize the label representations to the classification process. Specifically, considering a task f with class set $\mathcal{C}_f$, we build the weight matrices of the classifier from the label embeddings $\{\widehat{h}_c | c \in \mathcal{C}_f\}$. For the outputs $\widetilde{h}$ from the meta-model, the predicted probability $P(c)$ of class c is obtained as:

$$\widetilde{h}' = \sigma(W_1\widetilde{h} + b_1),\ \widehat{h}'_c = \sigma(W_2\widehat{h}_c + b_2), \tag{5}$$

$$s_c = W_3(\widetilde{h}' \odot \widehat{h}'_c) + b_3, \tag{6}$$

$$P(c) = \exp(s_c) / \textstyle\sum_{c\in\mathcal{C}_f} \exp(s_c), \tag{7}$$

where $\{W_i, b_i\}_{i=1}^3$ are learnable transformation matrices. The classifier is agnostic to the number and content of the labels, offering a universal interface for transferring from pre-training to downstream tasks.

3.3. Self-Supervised In-Context Learning

Self-Supervision Construction. To ensure the generalizability of our framework across various tasks, it is imperative to create prompt sequences that encompass a wide range of data and labels. While collecting data from multiple supervised datasets is a valid approach, the VL domain has limited tasks and label spaces compared to the language [4] or vision [36] domains. To address this issue, we propose creating data-label pairs in a self-supervised manner. Firstly, we establish a general multimodal label set by investigating *concepts* in unannotated image-text pairs. Specifically, we parse texts to identify salient spans that are typically pertinent to both the language and vision domains, such as nouns, verbs, adjectives, and adverbs. The spans are then collected as the pre-training label set $\mathcal{C}$. Next, inspired by the literature of unsupervised question answering [20, 50, 71], we create data for each label by *considering the label as the missing semantics for data*. In particular, for a given label $c \in \mathcal{C}$, we select image-text pairs containing c and replace the span of c with a mask token `[MASK]`. The mask token indicates missing information, and data queries for the same information can be grouped together. This approach is related to Masked Language Modeling [15], which aims to predict masked tokens from the context of

a sentence, but we concentrate on using the mask token to create homogeneous data for different labels. This design allows us to generate diverse data-label pairs in large quantities for constructing prompts that support generalization.

Demonstration Formation. Through the utilization of self-supervised data, an extensive amount of prompts can be employed for pre-training. However, we observe that the formation of demonstrations has a significant impact on the emergence of the in-context ability. Specifically, we identify that models tend to disregard the demonstrations and rely solely on the query data for predictions. We believe this is a result of the insufficient correlation between the query data and demonstrations, as models tend to learn from the readily accessible information regarding predictions [21, 57]. To this end, we propose three types of prompts: *label-in-demo (LID)*, *data-in-demo (DID)*, and *out-demo (OD)* prompts, to incentivize the model capacity from various perspectives. The LID and DID prompts aim to enhance the correlation between the query data and demonstrations, whereas the OD prompts aim to reinforce the utilization of query data. Specifically, given a query data x^q belonging to class c^q, we sample n classes $\mathcal{C}^d$ from the pre-training label set $\mathcal{C}$, and then sample an equal amount of data for each class to create the LID prompts: $\mathcal{X}^{lid} \subset \{g(c)|c \in (\mathcal{C}^d \cup \{c^q\})\}$, where $g : \mathcal{C} \rightarrow \mathcal{X}$ indicates the set of data with the same class. This approach simulates the few-shot learning regime [11, 62] and encourages model learning based on demonstrated information. While LID prompts consider the correlation on the label space, we propose to learn with correlated demonstrations also on the data space. This is achieved by retrieving similar data based on the vision-language representations to create the DID prompts: $\mathcal{X}^{did} \subset \textit{top-k}(sim(\mathcal{X}, x^q))$, where $sim(\cdot, \cdot)$ is a simple cosine similarity function. The LID and DID prompt jointly promote the models to utilize demonstrations for predictions. To balance the utilization of query data and demonstrations, the OD prompts are constructed by randomly sampling data from the datasets: $\mathcal{X}^{od} \subset \mathcal{X}$, which has relatively fewer benefits from demonstrations. We further introduce the *in-demo ratio* ρ to balance the exposure of OD or LID/DID prompts during pre-training. This enables us to further control the model's inclination to leverage demonstrations, which could vary across different tasks. The examples of LID/DID/OD prompts are shown in Fig. 4 and the learning process is summarized in Alg. 1

Correlation Embeddings. Utilizing the proposed prompts for learning allows the model to predict with variant reliance on the provided demonstrations. To enable the model to activate the corresponding capacity concerning the demonstrations, we introduce the *correlation embeddings* that specify the relations between the query data and demonstrations. Specifically, we add the embeddings to data representations of prompts, h^q and $\{\widehat{h}_i^d\}$, as follows:

$$h^q \leftarrow h^q + e_c,\ h_i^d \leftarrow h_i^d + e_c, \tag{8}$$

$$e_c = \sigma(W_c(\textstyle\sum_i (\widehat{h}_i^d \odot h^q)) + b_c), \tag{9}$$

where W_c and b_c are learnable parameters. This design provides controllability of predictions conditioned on demonstrations, benefiting tasks with diverse prompt distributions.

Algorithm 1: Self-Supervised In-Context Learning

```
Construct self-supervised dataset D = {(x, f(x))}.
Let X be the data set, C be the class set.
for (x^q, f(x^q)) ∈ D do
    if δ_1 < ρ, where δ_1 ~ U(0,1) then
        if δ_2 = 0, where δ_2 ~ B(0.5) then
            Sample classes C^d ⊂ C;
            Sample data X^d = X^lid ⊂ {g(c)|c ∈ (C^d ∪ {c^q})};
        else
            Sample data X^d = X^did ⊂ top-k(sim(X, x^q));
        end
    else
        Sample data X^d = X^out ⊂ X;
    end
    π = (X^d, f(X^d), x^q) = (x_1^d, f(x_1^d), ..., x^q);
    Compute loss L(f(x^q)|π, θ) from Eq. 3 and update θ;
end
```

4. Experiments

4.1. Experimental Settings

Pre-training Datasets. We construct the self-supervised dataset proposed in Sec. 3.3 from four image-text datasets, including COCO [9], Visual Genome [32], Conceptual Captions [55], and SBU Captions [46]. The labels are designed to encompass nouns, verbs, adjectives, and adverbs extracted from image-text pairs. The dataset is curated to contain 4 million data, and further expansion is feasible. To enhance learning efficiency, we preprocess the data representations offline.

Downstream Datasets. We benchmark SINC on various VL tasks, including multimodal fast concept binding [67], visual question answering (VQAv2 [22]), visual entailment (SNLI-VE [70]) and visual reasoning (NLVR^2 [61]). These tasks exhibit diverse data formats, enabling us to examine the properties of ICL across different scenarios.

Implementation Details. The meta-model is a 12-layer decoder-only transformer, and the multi-source feature fuser comprises a single cross-attention layer. For data representation, METER [18], ViT [17], and RoBERTa [39] are considered as the vision-language, vision, and language knowledge sources. During pre-training, we use 8 demonstrations. For DID prompts, we leverage Faiss [29] to retrieve related data based on VL representations. For LID prompts, we sample 1 class in addition to the query class.

The model is pre-trained for 500k steps with 4k warm-up steps. We monitor pre-training performance with LID and OD prompts from a separate validation set. For downstream tasks, we leverage DID prompts for ICL evaluation.

4.2. Comparison with Prior Arts

We first conduct experiments on the fast concept binding [67], which is established to evaluate models' ability to associate a word with a visual category in few-shot settings. Tab. 1 demonstrates that SINC significantly outperforms both ICL (row 3) and gradient-based (GD) methods (rows 1-2) by at least 57.1%/50.1%/51.5% under 2-/6-/10-shots. The main benefits of Frozen (row 3) come from the huge pre-trained LM, which also limits the model's capacity to adapt to new concepts from a few demonstrations since it depends mostly on pre-learned knowledge. In contrast, SINC possesses the advantage of reducing the dependency on linguistic cues or template designs for leveraging demonstrations. This property allows SINC to better adapt to novel tasks, thereby highlighting its superiority. Notably, GD methods (rows 1-2) hardly learn an effective predictor to tackle novel words, even METER-P reuses the language model head as in the prompt learning scheme [38].

Next, we evaluate SINC on various real-world VL tasks. Tab. 2 presents the comparisons on visual entailment and reasoning tasks under few-shot regimes, which have received limited attention in prior research. The results demonstrate that SINC is capable of tackling tasks that require reasoning skills. Notably, the data representations for NLVR^2 are obtained by combining two images, which are generally not seen during pre-training, highlighting SINC's ability to generalize to diverse data representations. Tab. 3 further presents the comparisons on visual question answering, indicating that SINC outperforms GD methods (rows 4-7) under the 4-shot setting and remains competitive for higher shot numbers. Notably, [58] (row 7) is tailored for the VQAv2 dataset. It uses a pre-trained language model to filter candidate answers and generate text templates for matching images with CLIP [48], and the computation cost would increase rapidly with the number of candidate answers. We emphasize that SINC confers further benefits in enabling predictions without any parameter updates. Additionally, our framework is generalizable across different tasks, thus eliminating the need for problem-specific tailoring as prior works. From the comparisons, we also note that the benefits from increasing shot number tend to plateau for SINC, which aligns with prior research [2]. Further investigation is presented in Sec. 4.3. Compared to previous ICL methods (rows 1-3), SINC employs a significantly lower number of learnable and frozen parameters (at least 3.5 and 13.2 times less) to achieve ICL. While we aim to compare models with a more manageable size (<1B), SINC can still outperform Frozen (>7B) substantially, emphasizing the merits of learning the ICL explicitly.

Model	GD	# of Params. Learn / Frozen	Fast Concept Binding 2- / 6- / 10-shot
METER-C [18]	✓	319M / -	50.00 / 50.43 / 50.98
METER-P [18]	✓	319M / -	50.00 / 50.33 / 51.20
Frozen [67]	✗	438M / 7B	53.40 / 57.90 / 58.90
SINC(ours)	✗	82M / 319M 124M / 529M	76.56 / 79.88 / 82.28 **83.88 / 86.92 / 89.24**

Table 1. Performance comparisons on fast concept binding. The "-C" and "-P" methods refer to initializing classifiers or reusing LM heads for predictions. GD specifies the need of gradient descent.

Model	GD	# of Params. Learn / Frozen	SNLI-VE dev / test	NLVR^2 dev / test
VL-T5 [13]	✓	224M / -	37.54/38.51	52.65/51.50
METER [18]	✓	319M / -	49.21/49.14	55.03/55.03
SINC(ours)	✗	82M / 319M 124M / 529M	53.03/53.02 **54.71/54.98**	55.27/55.30 **58.94/59.04**

Table 2. Performance comparisons on SNLI-VE and NLVR^2 with 16 demonstrations. GD specifies the need of gradient descent.

Model	GD	# of Params. Learn / Frozen	VQAv2 4- / 16- / 32-shot
Frozen parameter counts > 1B			
Frozen [67]	✗	438M / 7B	38.20 / - / -
PICa [76]	✗	- / 175B	- / 54.30 / -
Flamingo [2]	✗	10B / 70.5B	63.10 / 66.80 / 67.60
Frozen parameter counts < 1B			
METER [18]	✓	319M / -	23.53 / 24.43 / 26.89
VL-T5 [13]	✓	224M / -	- / 31.80 / -
FewVLM [28]	✓	224M / -	45.10 / 48.20 / -
TAP-C [58]	✓	0.3M / 229M	45.87 / **48.89** / **50.18**
SINC(ours)	✗	82M / 319M 124M / 529M	46.21 / 46.60 / 46.82 **47.21** / 48.25 / 48.58

Table 3. Performance comparisons on VQAv2. GD specifies the need of gradient descent. The second best scores are underlined.

4.3. Main Properties

Learning Dynamics. We monitored the performance of the model during pre-training to investigate the emergence of in-context ability in Fig. 5. The results suggest a trade-off between the performance of OD and LID prompts in two regards. Firstly, an increase in the in-demo ratio is associated with a higher LID performance and the early development of in-context ability. Notably, during validation, we mapped the labels of the LID prompts to random ones, precluding the possibility of accurate predictions based on memorization, and therefore establishes that the models indeed leverage information from the demonstrations. Secondly, we found that the performance of LID prompts would de-

crease as the training proceeds to later stages, while the performance of OD prompts continues to improve. We hypothesize that this trade-off could stem from the learning interference across data distributions [47, 66] since different prompts aim to acquire distinct and possibly opposing capacities. The alleviation of this trade-off is also worth exploring as a future research direction. Furthermore, we observed that the performance of LID prompts remains at binary chances for a 0.0 in-demo ratio. Overall, our findings suggest that models may not efficiently learn to predict with demonstration without specific incentivization, and our framework effectively addresses this issue through the learning with self-supervised prompts.

Different In-Demo Ratio. Fig. 6 presents the in-context benefits (ICB) achieved by models trained with different in-demo ratios across various tasks. Our results reveal that a higher in-demo ratio confers significant advantages for tasks such as fast concept binding, which entails simple classification but requires sufficient information from demonstrations. Moreover, we observe that VQAv2 attains a higher ICB compared to NLVR^2 and SNLI-VE. We attribute this finding to their smaller label space, which comprises only two and three categories, thus rendering the construction of the label space relatively easy and relying less on demonstrations. Notably, since the frozen VL model we used [18] is pre-trained with image-text matching (ITM), demonstrations may readily activate the binary classification ability, resulting in the closed ICB for NLVR^2 across different in-demo ratios. Importantly, we note that the peak of ICB varies across tasks, indicating that different tasks require varying degrees of in-context ability. We believe this highlights the necessity of providing controllability in leveraging demonstrations, which we primarily achieved through the design of correlation embeddings.

Different Number of Demonstrations. Fig. 7 shows the impact of varying numbers of demonstrations on the in-context benefit (ICB). Significantly, different tasks exhibit varying sensitivity to the number of demonstrations, with a notable performance saturation as the length of prompts increases. We hypothesize that the saturation may arise due to the generalization issues of Transformer models. Specifically, previous studies have highlighted significant generalization deficiencies in Transformers with respect to sequence length [3], and attention could be distracted on long sequences, resulting in degraded performance [60]. Thus, relevant techniques may be applied to mitigate this issue, which we identify as an important future research direction. Notably, the efficient learning scheme of SINC allows us to investigate such issues with higher shots to further understand the properties of ICB.

Order Sensitivity. To explore the influence of demonstration order on SINC, we evaluate standard deviations for performance on VQAv2/SNLI-VE/NLVR2, yielding respective values of 0.41/0.35/0.53 for the 0.2 in-demo ratio and 4 demonstrations, which is generally not significant compared to mean values. Our results indicate that SINC exhibits low sensitivity to the order of demonstrations. We attribute this property to the specialized learning process of SINC, which facilitates exposure to diverse prompts.

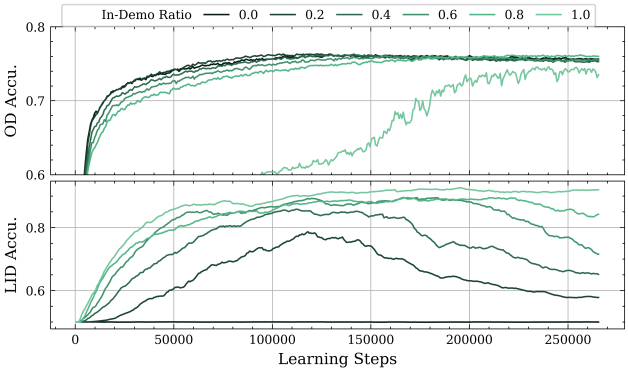

Figure 5. Dynamics of validation performance for OD and LID prompts during pre-training.

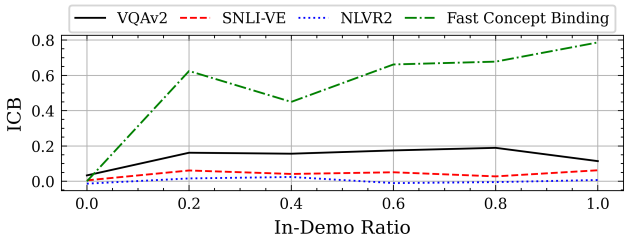

Figure 6. In-Context Benefit (ICB) for models learned in different in-demo ratios. The number of demonstrations is 2 and 4 for fast concept binding and other tasks, respectively.

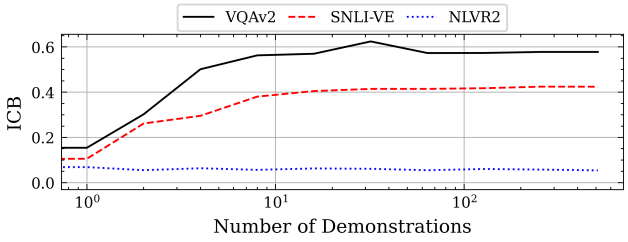

Figure 7. In-Context Benefit (ICB) for different numbers of demonstrations. The model is learned in a 0.2 in-demo ratio.

Different Settings. We evaluate the efficacy of SINC on various settings, as demonstrated in Tab. 4. Our results reveal that decreasing the meta-model size from 124M to 82M does not lead to a significant reduction in performance, while concurrently enhancing operational efficiency (row 2). Furthermore, increasing the size of the pre-training datasets yields improved performance, indicating the potential for further scalability (row 3). Our ablation analysis also demonstrates the performance benefits of utilizing multi-source data representation (rows 4-5). Remarkably,

even in the absence of pre-trained VL models, SINC could effectively learn for ICL by leveraging both vision and language representations (row 5). Moreover, our framework is highly flexible, enabling the use of different VL models, e.g., BLIP2 [34] and ViLT [31], for learning an ICL framework. The results demonstrate that SINC can generalize across various VL models, with feature quality having a further impact on performance (rows 6-7). Therefore, it allows us to boost performance with stronger feature extraction models. Additionally, we explored the impact of demonstration selection, where OD prompts were used instead of LID prompts for evaluation (row 8). The results indicate that incorporating relevant demonstrations can further boost performance, aligning with prior studies [37, 79].

4.4. Learning Efficiency

SINC decouples the learning of in-context ability by introducing a meta-model that operates on the representations generated by frozen models. By doing so, the frozen models do not participate in the backward process, thereby reducing the computational cost significantly. To evaluate the effectiveness of this scheme, we conduct experiments by comparing the learning cost, including GFLOPs and memory footprint of SINC with that of the Frozen [67], under comparable settings. More details are available in Appendix. The results, presented in Tab. 5, clearly indicate that SINC achieves a significant reduction in learning cost compared to Frozen. This is a desirable property, especially considering the rapid evolution of pre-trained models, as it allows us to efficiently retrain the meta-model and leverage the latest knowledge sources. Additionally, Tab. 5 shows the inference cost of the representation obtained from frozen pre-trained models in bottom rows. Notably, our proposed architecture enables flexible selections for feature extractors, providing superior scalability and generalizability.

5. Discussion and Conclusion

Current ICL techniques in the vision-language (VL) domain are heavily reliant on large pre-trained language models, which may hinder their scalability and applicability. In this paper, we identify that this dependence arises from the ambiguous objective of acquiring ICL. To this end, we propose a novel framework, named SINC, that decouples the acquisition of ICL from VL pre-training and incentivizes it from both architectural and data perspectives. Our proposed method not only achieves superior performance compared to previous methods but also has a lower learning cost, making it advantageous for leveraging pre-trained models in a black-box setting. This property is particularly crucial for models with inaccessible parameters, such as ChatGPT. Moreover, SINC provides a general interface for exploring ICL in real-world applications and uncovering properties that can be utilized in future studies. We envision various directions for future research based on the proposed framework, e.g., alleviating the trade-off for learning with different prompts, enabling controllability conditioned on the given demonstrations, and facilitating generalization for higher shot numbers. We hope that our framework and the perspectives provided in this paper will further drive the development of ICL methods in the VL domain.

Setting	VQAv2 val	SNLI-VE dev / test	NLVR2 dev / test
Default	44.42	53.35/53.23	54.97/56.39
Scale of meta-model			
124M → 82M	42.67 -1.75	51.89/52.00 -1.35	53.25/54.44 -1.84
Scale of pre-training dataset			
4M → 8M	45.22 +0.80	53.40/54.00 +0.41	55.02/56.60 +0.13
Sources of data representations			
All → VL	43.87 -0.55	52.89/53.00 -0.35	53.55/55.31 -1.25
All → V + L	36.14 -8.28	51.68/51.28 -1.81	52.90/52.67 -2.90
Different VL Models			
[18] → [34]	41.78 -2.64	53.56/53.65 +0.32	57.03/56.85 +1.26
[18] → [31]	37.46 -6.96	45.50/45.06 -8.01	52.84/52.94 -2.79
Demonstration Selection			
w/ → w/o	42.80 -1.62	52.30/52.05 -1.12	52.46/53.39 -2.76

Table 4. Different settings of SINC. Across settings, the in-demo ratio is 0.2 and the number of demonstrations is 4 for fair comparisons, which may not yield optimal values for tasks.

Scale	Input Size	Learnable Params.	Computation (GFLOPs)	Memory (GB)
Frozen [67]				
Small	1	86.39M	1898.80	28.44
Large	1	304.35M	2162.63	38.79
SINC (ours)				
Small	1	81.91M	2.15	2.18
Large	1	354.82M	14.67	14.93
- VL Feats.	9	-	301.06	3.26
- V Feats.	9	-	281.05	3.06
- L Feats.	9	-	792.95	11.45

Table 5. Learning efficiency comparisons of SINC with [67]. The cost includes both forward and backward processes if required. We report the cost of feature inference with an input size of 9 for 8 demonstrations and 1 query data.

Acknowledgement

This work was supported in part by the National Science and Technology Council of Taiwan under Grants NSTC-109-2221-E-009-114-MY3 and NSTC-112-2221-E-A49-094-MY3.

References

[1] Ekin Akyürek, Jacob Andreas, Dale Schuurmans, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *International Conference on Learning Representations*, pages 1–8, 2023. 2, 3

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 1–8, 2022. 1, 2, 3, 7

[3] Cem Anil, Yuhuai Wu, Anders Johan Andreassen, Aitor Lewkowycz, Vedant Misra, Vinay Venkatesh Ramasesh, Ambrose Slone, Guy Gur-Ari, Ethan Dyer, and Behnam Neyshabur. Exploring length generalization in large language models. In *Advances in Neural Information Processing Systems*, pages 1–8, 2022. 8

[4] Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q. Tran, Dara Bahri, Jianmo Ni, Jai Gupta, Kai Hui, Sebastian Ruder, and Donald Metzler. Ext5: Towards extreme multi-task scaling for transfer learning. In *International Conference on Learning Representations*, pages 1–8, 2022. 5

[5] Hritik Bansal, Karthik Gopalakrishnan, Saket Dingliwal, Sravan Bodapati, Katrin Kirchhoff, and Dan Roth. Rethinking the role of scale for in-context learning: An interpretability-based case study at 66 billion scale. *arXiv preprint arXiv:2212.09095*, pages 1–8, 2022. 3

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, pages 1877–1901, 2020. 1, 3, 4

[7] Stephanie C.Y. Chan, Adam Santoro, Andrew Kyle Lampinen, Jane X Wang, Aaditya K Singh, Pierre Harvey Richemond, James McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. In *Advances in Neural Information Processing Systems*, pages 1–8, 2022. 2, 3

[8] Mingda Chen, Jingfei Du, Ramakanth Pasunuru, Todor Mihaylov, Srini Iyer, Veselin Stoyanov, and Zornitsa Kozareva. Improving in-context few-shot learning via self-supervised training. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3558–3573, 2022. 3

[9] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, pages 1–8, 2015. 1, 6

[10] Yanda Chen, Ruiqi Zhong, Sheng Zha, George Karypis, and He He. Meta-learning via language model in-context tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 719–730, 2022. 3

[11] Yi-Syuan Chen and Hong-Han Shuai. Meta-transfer learning for low-resource abstractive summarization. *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12692–12700, 2021. 2, 3, 6

[12] Yi-Syuan Chen, Yun-Zhu Song, and Hong-Han Shuai. Spec: Summary preference decomposition for low-resource abstractive summarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pages 603–618, 2023. 3

[13] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *Proceedings of the 38th International Conference on Machine Learning*, pages 1931–1942, 2021. 7

[14] Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Zhifang Sui, and Furu Wei. Why can gpt learn in-context? language models secretly perform gradient descent as meta optimizers. *arXiv preprint arXiv:2212.10559*, pages 1–8, 2022. 2

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019. 5

[16] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*, pages 1–8, 2022. 3

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, pages 1–8, 2021. 6

[18] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, and Michael Zeng. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18166–18176, 2022. 6, 7, 8, 9

[19] Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. In *Advances in Neural Information Processing Systems*, pages 1–8, 2022. 1, 3

[20] Michael Glass, Alfio Gliozzo, Rishav Chakravarti, Anthony Ferritto, Lin Pan, G P Shrivatsa Bhargav, Dinesh Garg, and Avi Sil. Span selection pre-training for question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2773–2782, 2020. 2, 5

[21] Tanya Goyal, Jiacheng Xu, Junyi Jessy Li, and Greg Durrett. Training dynamics for text summarization models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2061–2073, 2022. 6

[22] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2017. 1, 6

[23] Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051, 2021. 3

[24] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12976–12985, 2021. 3

[25] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, pages 1–8, 2020. 3

[26] Yatai Ji, Rongcheng Tu, Jie Jiang, Weijie Kong, Chengfei Cai, Wenzhe Zhao, Hongfa Wang, Yujiu Yang, and Wei Liu. Seeing what you miss: Vision-language pre-training with semantic completion learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6789–6798, 2023. 3

[27] Bernal Jimenez Gutierrez, Nikolas McNeal, Clayton Washington, You Chen, Lang Li, Huan Sun, and Yu Su. Thinking about GPT-3 in-context learning for biomedical IE? think again. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4497–4512, 2022. 1

[28] Woojeong Jin, Yu Cheng, Yelong Shen, Weizhu Chen, and Xiang Ren. A good prompt is worth millions of parameters: Low-resource prompt-based learning for vision-language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2763–2775, 2022. 3, 7

[29] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, pages 535–547, 2019. 6

[30] Boseop Kim, HyoungSeok Kim, Sang-Woo Lee, Gichang Lee, Donghyun Kwak, Jeon Dong Hyeon, Sunghyun Park, Sungju Kim, Seonhoon Kim, Dongpil Seo, Heungsub Lee, Minyoung Jeong, Sungjae Lee, Minsub Kim, Suk Hyun Ko, Seokhun Kim, Taeyong Park, Jinuk Kim, Soyoung Kang, Na-Hyeon Ryu, Kang Min Yoo, Minsuk Chang, Soobin Suh, Sookyo In, Jinseong Park, Kyungduk Kim, Hiun Kim, Jisu Jeong, Yong Goo Yeo, Donghoon Ham, Dongju Park, Min Young Lee, Jaewook Kang, Inho Kang, Jung-Woo Ha, Woomyoung Park, and Nako Sung. What changes can large-scale language models bring? intensive study on HyperCLOVA: Billions-scale Korean generative pretrained transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3405–3424, 2021. 3

[31] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 5583–5594, 2021. 9

[32] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, pages 32–73, 2017. 6

[33] Renaud Lambiotte, Lionel Tabourier, and Jean-Charles Delvenne. Burstiness and spreading on temporal networks. *The European Physical Journal B*, pages 1–4, 2013. 2, 3

[34] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, pages 1–8, 2023. 3, 9

[35] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems*, pages 9694–9705, 2021. 3

[36] Hanxue Liang, Zhiwen Fan, Rishov Sarkar, Ziyu Jiang, Tianlong Chen, Kai Zou, Yu Cheng, Cong Hao, and Zhangyang Wang. M³vit: Mixture-of-experts vision transformer for efficient multi-task learning with model-accelerator co-design. In *Advances in Neural Information Processing Systems*, pages 1–8, 2022. 5

[37] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, 2022. 2, 3, 9

[38] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, pages 1–35, 2023. 7

[39] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, pages 1–8, 2019. 6

[40] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, 2022. 1, 2, 3

[41] Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Noisy channel language model prompting for few-shot text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5316–5330, 2022. 3

[42] Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. MetaICL: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2791–2809, 2022. 3

[43] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, pages 1–8, 2022. 2, 3

[44] Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. Reframing instructional prompts to GPTk's language. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 589–612, 2022. 3

[45] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, pages 1–8, 2022. 3

[46] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, pages 1–8, 2011. 6

[47] Jonathan Pilault, Amine El hattami, and Christopher Pal. Conditionally adaptive multi-task learning: Improving transfer learning in NLP using fewer parameters & less data. In *International Conference on Learning Representations*, pages 1–8, 2021. 8

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763, 2021. 7

[49] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, pages 1–8, 2019. 4

[50] Ori Ram, Yuval Kirstain, Jonathan Berant, Amir Globerson, and Omer Levy. Few-shot question answering by pretraining span selection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3066–3079, 2021. 2, 5

[51] Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854, 2022. 3

[52] Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, 2022. 1, 3

[53] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, 2016. 2

[54] M Ángeles Serrano, Alessandro Flammini, and Filippo Menczer. Modeling statistical properties of written text. *PloS one*, page e5372, 2009. 2, 3

[55] 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)*, pages 2556–2565, 2018. 6

[56] Seongjin Shin, Sang-Woo Lee, Hwijeen Ahn, Sungdong Kim, HyoungSeok Kim, Boseop Kim, Kyunghyun Cho, Gichang Lee, Woomyoung Park, Jung-Woo Ha, and Nako Sung. On the effect of pretraining corpora on in-context learning by a large-scale language model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5168–5186, 2022. 1, 2, 3

[57] Sahil Singla and Soheil Feizi. Salient imagenet: How to discover spurious features in deep learning? In *International Conference on Learning Representations*, pages 1–8, 2022. 6

[58] Haoyu Song, Li Dong, Weinan Zhang, Ting Liu, and Furu Wei. CLIP models are few-shot learners: Empirical studies on VQA and visual entailment. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6088–6100, 2022. 3, 7

[59] Yisheng Song, Ting Wang, Puyu Cai, Subrota K Mondal, and Jyoti Prakash Sahoo. A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. *ACM Comput. Surv.*, pages 1–18, 2023. 3

[60] Yun-Zhu Song, Yi-Syuan Chen, and Hong-Han Shuai. Improving multi-document summarization through referenced flexible extraction with credit-awareness. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1667–1681, 2022. 8

[61] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, 2019. 3, 6

[62] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2019. 2, 6

[63] Yuchong Sun, Hongwei Xue, Ruihua Song, Bei Liu, Huan Yang, and Jianlong Fu. Long-form video-language pre-training with multimodal temporal contrastive learning. In

Advances in Neural Information Processing Systems, pages 38032–38045, 2022. 3

[64] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. LST: Ladder side-tuning for parameter and memory efficient transfer learning. In *Advances in Neural Information Processing Systems*, pages 1–8, 2022. 2, 3

[65] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5227–5237, 2022. 3

[66] Eleni Triantafillou, Hugo Larochelle, Richard Zemel, and Vincent Dumoulin. Learning a universal template for few-shot dataset generalization. In *Proceedings of the 38th International Conference on Machine Learning*, pages 10424–10433, 2021. 8

[67] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. In *Advances in Neural Information Processing Systems*, pages 200–212, 2021. 1, 2, 3, 4, 6, 7, 9

[68] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 1–8, 2017. 4

[69] Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. *arXiv preprint arXiv:2212.07677*, pages 1–8, 2022. 3

[70] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, pages 1–8, 2019. 3, 6

[71] Yumo Xu and Mirella Lapata. Generating query focused summaries from query-free resources. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6096–6109, 2021. 2, 5

[72] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5036–5045, 2022. 3

[73] Hongwei Xue, Yupan Huang, Bei Liu, Houwen Peng, Jianlong Fu, Houqiang Li, and Jiebo Luo. Probing intermodality: Visual parsing with self-attention for vision-and-language pre-training. In *Advances in Neural Information Processing Systems*, pages 4514–4528, 2021. 3

[74] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. CLIP-vip: Adapting pre-trained image-text model to video-language alignment. In *The Eleventh International Conference on Learning Representations*, pages 1–8, 2023. 3

[75] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15671–15680, 2022. 3

[76] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *AAAI*, pages 1–8, 2022. 3, 7

[77] Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, and Taeuk Kim. Ground-truth labels matter: A deeper look into input-label demonstrations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2422–2437, 2022. 2, 3

[78] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, pages 17283–17297, 2020. 2

[79] Yiming Zhang, Shi Feng, and Chenhao Tan. Active example selection for in-context learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9134–9148, 2022. 2, 9

[80] Zhengkun Zhang, Wenya Guo, Xiaojun Meng, Yasheng Wang, Yadao Wang, Xin Jiang, Qun Liu, and Zhenglu Yang. HyperPELT: Unified parameter-efficient language model tuning for both language and vision-and-language tasks. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11442–11453, 2023. 3

[81] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, pages 12697–12706, 2021. 2, 3

[82] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 2