# ワールド・トゥ・ワード ビジョン言語モデルにおける高速マッピングによるオープン語彙の獲得

Ziqiao Ma Joyie Pan Joyce Chai ミシガン大学コンピューターサイエンス・エンジニアリング学部
{marstin,jiayipan,chaijy}@umich.edu

## Abstract

言語単位を物理世界における参照語(グラウンディングと呼ばれる)に接続する能力は、単語の根拠となる意味を学習し理解する上で極めて重要である。人間は新しい単語学習において高速なマッピングを示すが、現代の視覚言語モデルが本当に言語を根拠のある意味で表現できるのか、また、根拠が新しい単語学習をさらに促進するのかは不明なままである。そこで、我々はGrounded Open Vocabulary Acquisition (GOVA)を導入し、オープンワールド言語学習におけるグラウンディングとブートストラップについて検討する。最初の試みとして、我々はWorld-to-Words (W2W)を提案する。W2Wは、目的語として接地を強調する画像-テキストペアで事前学習を行うことで、視覚的に根拠のある新しい言語モデルである。W2Wはより首尾一貫した高速な接地語学習であり、事前学習で獲得した接地能力により、モデルが未見語をより迅速かつ頑健に学習できることを示す(^{1})。

## 1 Introduction

言語は、物理世界での感覚運動経験を通じて学習される(Bisk et al.、2020)。言語ユニットを物理世界での参照元に接続する能力(グラウンディングと呼ばれる)は、単語の根拠となる意味の学習と理解に重要な役割を果たす(Harnad, 1990)。図1に示すように、人間の読者は、名詞句を画像で捉えられた対応する実体に容易に接地することができる。人間の学習者が「インシネーター」という言葉を初めて使う場合でも、言語や視覚的文脈を通して対象物を見つけ出し、その意味を獲得することができるのである。実際、ファストマッピングと呼ばれる最小限の情報だけで新しい単語学習をブートストラップするこの能力は、認知的に豊富に実証されている。

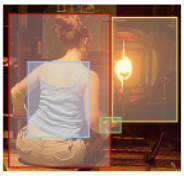



図1:「インシネーター」(黄色でハイライト)という用語が人間の学習者にとって新しいものである場合でも、彼らは接地によって知覚された世界の中で最も可能性の高い参照元(黄色のバウンディングボックスで示される)を見つけることができる。

人間の言語習得に関する文献(Carey and Bartlett, 1978; Carey, 1978; Golinkoff et al., 2000; Smith and Yu, 2008)。最近、視覚言語モデル(VLM)の事前学習にかなりの努力が払われている(Du et al.、2022a)。これらのモデルは、様々な下流の視覚・言語タスクにおいてエキサイティングな性能を発揮しているにもかかわらず、これらのモデルが、知覚された世界において根拠のある意味を持つ言語を本当に理解し、生成できるかどうか、また、根拠が新しい単語学習をさらにどのように打ち砕くことができるかは依然として不明である。これらの疑問は、科学的な観点からも工学的な観点からも興味深い。科学的な観点からは、言語学習者にとって根拠は極めて重要であり、子どもは発話を生成し(Tanenhaus et al., 1995; Meyer et al., 1998)、理解する(Smith et al., 2007)際に環境中の意図するオブジェクトに注目するようになる。工学的な観点からは、グラウンデッドビジョン言語データセット(細かい単語オブジェクトマッピングを持つ画像-テキストペア)が利用可能であっても, 2015)、コストのかかる接地アノテーションは、学習時間中に語彙空間全体をほとんどカバーすることができない。事前学習されたモデルを基に、エージェントは単語とオブジェクトのマッピングを行わずに、数ショットの画像とテキストのペアで接地した新しい単語を学習する能力を持つことが重要である。このため、我々は、オープンワールド言語学習における接地とブートストラップを検討するためのスケーラブルな定式化であるGrounded Open Vocabulary Acquisition (GOVA)を紹介する。この定式化では、lan-.

*Equal contribution.
[1]Code available at https://github.com/sled-group/world-to-words.



計算言語学会第61回年次大会講演論文集
*Volume 1: Long Papers*, pages 524–544
July 9-14, 2023 

# World-to-Words: Grounded Open Vocabulary Acquisition through Fast Mapping in Vision-Language Models

**Ziqiao Ma*** **Jiayi Pan*** **Joyce Chai**
Computer Science and Engineering Division, University of Michigan
{marstin,jiayipan,chaijy}@umich.edu

## Abstract

The ability to connect language units to their referents in the physical world, referred to as *grounding*, is crucial to learning and understanding grounded meanings of words. While humans demonstrate fast mapping in new word learning, it remains unclear whether modern vision-language models can truly represent language with their grounded meanings, and how grounding may further bootstrap new word learning. To this end, we introduce Grounded Open Vocabulary Acquisition (GOVA) to examine grounding and bootstrapping in open-world language learning. As an initial attempt, we propose World-to-Words (W2W), a novel visually-grounded language model by pre-training on image-text pairs highlighting grounding as an objective. Through extensive experiments and analysis, we demonstrate that W2W is a more coherent and fast grounded word learner, and that the grounding ability acquired during pre-training helps the model to learn unseen words more rapidly and robustly.[1]

## 1 Introduction

Language is learned through sensorimotor experience in the physical world (Bisk et al., 2020). The ability to connect language units to their referents in the physical world, referred to as *grounding*, plays an important role in learning and understanding grounded meanings of words (Harnad, 1990). As shown in Figure 1, a human reader would easily ground noun phrases to the corresponding entities captured in the image. Even when the term "incinerator" is new to human learners, they can still locate the object of interest through the language and visual context, and acquire its meaning. In fact, this ability to bootstrap new word learning with only minimal information, known as *fast mapping*, is demonstrated abundantly in cognitive literature on human language acquisition (Carey and Bartlett, 1978; Carey, 1978; Golinkoff et al., 2000; Smith and Yu, 2008).

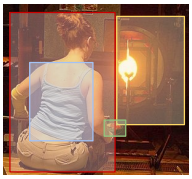



Figure 1: Even when the term "incinerator" (highlighted yellow) is new to human learners, they can still locate the most likely referent (indicated by the yellow bounding box) in the perceived world by grounding.

Recently, there has been a substantial effort on pre-training vision-language models (VLMs) (Du et al., 2022a). Despite the exciting performance of these models on a variety of downstream vision and language tasks, it remains unclear whether these models can truly understand or produce language with their grounded meanings in the perceived world, and how grounding may further bootstrap new word learning. These questions are of interest from both a scientific and an engineering point of view. From a scientific perspective, grounding is crucial to language learners, as children attend to intended objects in the environment when producing (Tanenhaus et al., 1995; Meyer et al., 1998) and comprehending (Smith et al., 2007) utterances. From an engineering perspective, even with the availability of grounded vision language datasets (image-text pairs with fine-grained word-object mappings) (Plummer et al., 2015), the costly grounding annotation can hardly cover the whole vocabulary space during the training time. Building upon the pre-trained models, it's important for the agent to have the ability to learn grounded new words in a few shots of raw image-text pairs without word-object mappings.

To this end, we introduce Grounded Open Vocabulary Acquisition (GOVA), a scalable formulation to examine grounding and bootstrapping in open-world language learning. In this formulation, lan-

*Equal contribution.
[1]Code available at https://github.com/sled-group/world-to-words.



*Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*
*Volume 1: Long Papers*, pages 524–544
July 9-14, 2023 

ージ学習とは、言語的な文脈で単語を予測する学習と、その単語を物理的な世界に接地する学習の組み合わせである。この定式化のもと、我々はモデルがまず事前学習で接地能力を獲得し、次にこの能力を教師なしで未見の単語を学習するために転移する枠組みを探索する。最初のステップとして、我々は最近の検出変換器(DETR)(Carion et al., 2020; Kamath et al., 2021)の進歩に動機づけられた新しい視覚的根拠言語モデルであるWorld-to-Words(W2W)を開発した。既存の多くのVLMと比較して、W2Wは明示的な物体表現に対して言語モデリングを行う。このモデルは、まず事前学習で接地能力を獲得し、接地された監視が利用できなくなったときに、この固有の能力を未知の単語の学習に移行させる。我々の実証結果は、単語とその参照元を対応付ける学習が、接地された単語の獲得に重要な役割を果たすことを示している。W2Wは、きめ細かい単語-物体マッピングを用いた事前学習により、単語の接地された意味を、見た目も見た目も、学習する上でより強い性能を示すが、他の競合するVLMベースラインと比較して桁違いに少ないデータしか得られない。この事前学習済みモデルは、さらに、少数の例で新しい接地語を効率的に学習するための基盤を提供することができる。さらに、単語学習におけるW2Wの潜在的な予測因子を理解するための詳細な分析を行い、人間の言語学習と比較して興味深い挙動を示す。この結果は、オープンワールドにおける接地言語学習に関する今後の研究の足がかりとなるであろう。

## 2 Grounded Open Vocabulary Acquisition (GOVA)

まず、GVA(Grounded Open Vocabulary Acquisition)タスクの定式化の重要な構成要素である、グラウンデッドワード獲得と新しい単語の数撃ちゃ当たる学習の設定を紹介することから始める。さらに、統一的な評価プロトコルを提示し、この問題のためにキュレーションしたデータセットを紹介する。

### 2.1 Grounded Word Acquisition

視覚言語タスクは、視覚的質問応答、視覚的コモンセンス推論など、過去に数多く開発されてきた。しかし、これらのタスクは主に、単語が対応する視覚的実体に根拠づけられているかどうかを精査することなく、タスクの終了性能に焦点を当てている。我々は、視覚言語モデルが、特に言語モデリングと物体定位の両方を通じて、単語の根拠ある意味を獲得する能力を有するかどうかを直接的に検証する定式化を検討する。

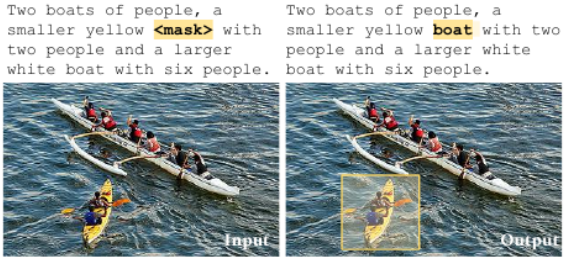



図2:単語接地タスクの一例。モデルは欠落した単語ボートを予測し、対応する小さい黄色のボートを画像中のまとまった位置に特定するようタスクが課される。

図2に単語獲得タスクの一例を示す。モデルには画像x $img_{\in I}$と不完全なキャプションx $cap_{\in T}$が提示され、その根拠となる単語w(例えば、名詞や形容詞)の1つがMASKに置き換えられている。このモデルは、利用可能なすべての文脈に基づいてこの欠落した単語$w_{\in V}$を予測し、対応するオブジェクトO w =_{{o} 1, o 2,-{--}, o n_{}}を画像中のバウンディングボックスを提案することによってローカライズするタスクが課せられている。全体として、接地語獲得タスクを解くことができるモデルは、関数 f :$_{I\times T\to V\times}$ $R^{4n}$. 言語モデリング部分は、オープンボキャブラリー単語を予測するクローズテストの形をとり、事前に学習した言語モデルを評価するために広く採用されている(Paperno et al.2016; Petroni et al.2019; Jin et al.2020)。しかし、言語モデリングだけでは、言語接地に関する包括的な評価を行うことができない。例えば、図2において、モデルは「ボート」という単語を正しく生成するが、誤ってその証拠を画像中の大きな白いボートに帰属させることがある。この制限に対処するために、我々はモデルが画像中の対応するオブジェクトをローカライズすることを要求する。この設計は、オブジェクト検出をオブジェクトローカライゼーションとクラス認識に分離することから動機づけられている(Singh et al., 2018; Zareian et al., 2021; Zhong et al., 2022)。これにより、ビジョンモデルは、あらかじめ定義された物体クラスの集合に依存することなく、物体らしさの感覚を発展させることができ、それによって、未知の物体への汎化を可能にする可能性がある。さらに、関連するタスクの設定との比較はセクション5で行われ、付録の図8に示されている。

### 2.2 Evaluation Metric

言語モデル評価において、性能評価によく用いられる指標は、標準的なヒット率-at-k (HR@k) 指標

guage learning is a combination of learning to predict a word in a linguistic context as well as learning to ground the word in the physical world. Under this formulation, we explore the framework in which the model first acquires the grounding ability during pre-training, and then transfers this ability to learn unseen words without grounding supervision. As an initial step, we developed World-to-Words (W2W), a novel visually grounded language model motivated by recent advances in detection transformers (DETR) (Carion et al., 2020; Kamath et al., 2021). Compared to many existing VLMs, W2W performs language modeling upon explicit object representations. The model first acquires the ability to ground during pre-training, and then transfers this intrinsic ability to learn unseen words when grounded supervision is no longer available.

Our empirical results show that learning to map words to their referents plays a significant role in grounded word acquisition. By pre-training with fine-grained word-object mappings, W2W demonstrates stronger performance in learning grounded meanings of words, both seen and unseen, yet with orders of magnitude fewer data compared to other competitive VLM baselines. The pre-trained model can further provide a foundation for efficient learning of new grounded words with a few examples. We further present an in-depth analysis to understand potential predictors of W2W in word learning, which demonstrates intriguing behaviors in comparison to human language learning. Our findings will provide a stepping stone for future work on grounded language learning in an open world.

## 2 Grounded Open Vocabulary Acquisition (GOVA)

We start by introducing the settings of *grounded word acquisition* and *few-shot learning of new words* tasks, which are two key components of the Grounded Open Vocabulary Acquisition (GOVA) task formulation. We further present a unified evaluation protocol and introduce the dataset we curated for this problem.

### 2.1 Grounded Word Acquisition

Many vision-language tasks have been developed in the past, *e.g.*, visual question answering, visual commonsense reasoning, etc. However, these tasks are mainly focused on the end task performance without scrutinizing whether words are grounded to their corresponding visual entities. We consider a formulation that directly examines if vision-language models have the ability to acquire grounded meanings of words, specifically, through both *language modeling* and *object localization*. Figure 2 shows an instance of the word acquisition task. A model is presented with an image $x_{\text{img}} \in \mathcal{I}$ and an incomplete caption $x_{\text{cap}} \in \mathcal{T}$ with one of its groundable words $w$ (*e.g.*, nouns and adjectives) replaced by a MASK. The model is tasked to predict this missing word $w \in \mathcal{V}$ based on all available context and localize the corresponding objects $O_w = \{o_1, o_2, \cdots, o_n\}$ in the image by proposing the bounding boxes of them. Overall, a model capable of solving the grounded word acquisition task is a function $f : \mathcal{I} \times \mathcal{T} \rightarrow \mathcal{V} \times \mathbb{R}^{4n}$.

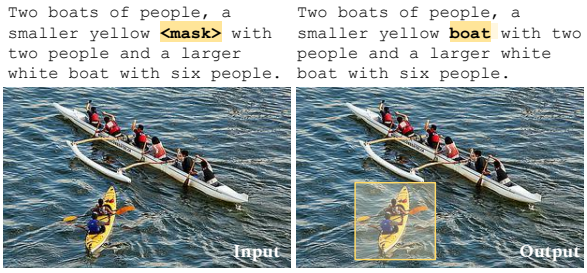



Figure 2: An instance of the word grounding task. Models are tasked to predict the missing word boat and localize the corresponding smaller yellow boat in the image coherently.

The language modeling part takes the form of a cloze test, which predicts an open vocabulary word and is widely adopted to evaluate pre-trained language models (Paperno et al., 2016; Petroni et al., 2019; Jin et al., 2020). However, language modeling alone fails to provide a comprehensive evaluation of language grounding. For example in Figure 2, a model may correctly produce the word "boat," but mistakenly attributes the evidence to the larger white boat in the image. To address this limitation, we require models to localize the corresponding object in the image. This design is motivated by the disentanglement of object detection into object localization and class recognition (Singh et al., 2018; Zareian et al., 2021; Zhong et al., 2022). It enables vision models to develop a sense of objectness without relying on a predefined set of object classes, thereby potentially allowing them to generalize to unseen objects. Further comparison with related task setups is discussed in Section 5 and illustrated in Figure 8 in the Appendix.

### 2.2 Evaluation Metric

In language model evaluation, the commonly used measures for assessing performance are the stan-

とパープレキシティである（Salazar et al., 2020; Jin et al., 2020)。マスク言語モデリングでは、単語wの対数パープレキシティは対数擬似パープレキシティとして定義される。

$$\log \mathrm{PPL}(w) = -\log P(w|x_{\mathrm{img}}, x_{\mathrm{cap}}) \quad (1)$$

物体検出の評価では、特に複数の参照語が可能なフレーズ接地について(Kamath et al.、2021)、Any-ProtocolとAll-Protocolが一般的に採用されている。n個の基底真理値バウンディングボックスB = b 1, b 2,_ _, b n_{}} とm個の予測バウンディングボックスB = b, b, b, bは交差-過剰結合であると仮定する。(IoU) e^{{} in e both e⁻ protocols⁻ f^{},} は次のように定義される。

$$\mathrm{IoU}_{\mathrm{any}} = \frac{1}{n} \sum_{i \in \{1,2,\cdots,n\}} \max_{j \in \{1,2,\cdots,m\}} \mathrm{IoU}(b_i, \widetilde{b_j}) \quad (2)$$

$$\mathrm{IoU}_{\mathrm{all}} = \mathrm{IoU}(\cup B, \cup \widetilde{B}) \quad (3)$$

しかし、これらのメトリクスは、クロスモーダルマッピングの正しさを考慮することなく、eのユニモーダル性能しか捉えていない。我々は、言語と視覚の性能を組み合わせるために、2つの新しいメトリクスを設計する。

- Grounded hit-rate (G-HR@k)、マスクされた単語が上位k個の候補に出現し、局在化IoUが0.5を超えるテストの割合。- Grounded perplexity (G-PPL)は以下の通り。

$$\log \text{G-PPL}(w) = \begin{cases} \text{IoU = 0 log PPL(w)\_ log IoU else} \end{cases} \quad (4)$$

## 2.3 新しい単語の少数ショット学習

接地データセット、すなわち、単語-オブジェクトマッピングアノテーションを持つ画像-テキストペアは利用可能であるが(Plummer et al., 2015)、大規模で、語彙空間全体をカバーするこのような細かいアノテーションを得ることは非現実的である_{V. }}。そこで、特に漸進的クラス学習の設定下で、少数ショット学習問題として根拠ある新しい単語学習を探求する(Mandziuk and Shastri, 1999; Kemker et al.、2018)。図 3 に、数発の新しい単語学習フレームワークの直感的な説明を示す。このフレームワークの下では、2つのステージで計算モデルが開発される。事前学習段階では、モデルは画像とキャプションのペアを受け取り、基本単語の集合$_V$ seen$_{\subseteq V}$. に対してきめ細かい単語とオブジェクトのアノテーションが行われる。事前学習後、モデルは少数の生のテキスト-画像ペアのサンプルを与えられ、それぞれがモデルが取得しなければならない未見単語の集合$_V$ unseen$_{\subseteq V}$を含んでいる。

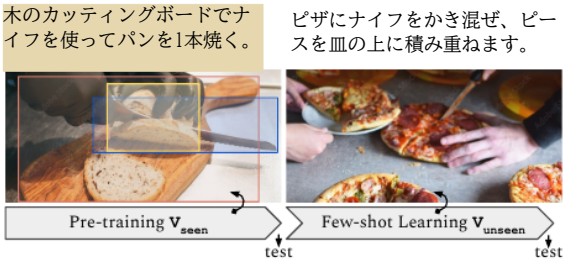



図3:数撃ちゃ当たるの新しい単語学習パラダイムの説明図。このモデルはまず、基本単語の集合($V_{seen}$)を持つ接地データセットで事前学習を行い、次に、少数の生のテキスト-画像ペアで未見単語の集合($V_{unseen}$)を取得しようとするものである。テストは各トレーニングセッションの後に行われる。

テストは各トレーニングステージの後に実行される。例えば、モデルは事前に学習された言語モデルで初期化された言語エンコーダでこれらの単語に遭遇したことがあるかもしれない。我々は、モデルがこれらの単語を参照語と対になって見ることがない、すなわち、単語の根拠となる意味が不明であるため、これらを「未見」と見なす。

## 2.4 Dataset Curation

我々は、Flickr30K Entitiesデータセット(Plummer et al., 2015)に基づいてデータセットを構築し、接地可能なフレーズとオブジェクトのバウンディングボックスの間の密な注釈を持つ画像-テキストペアを含んでいる。接地可能なフレーズと領域は、データセットによって、オブジェクトのバウンディングボックスを参照するテキストのチャンクとして定義される。単語接地インスタンスを構築するために、Stanza (Qi et al., 2020) を用いてキャプションを解析し、接地可能なフレーズ内の全ての単語を列挙し、NOUNまたはADJの品詞タグを持つものを特定する。これらの接地可能な単語は一度に1つずつMASKに置き換えられ、対応するバウンディングボックスにマッチングされる。データセットは、事前学習セット、未見語トレーニングセット、見語テストセット、未見語テストセットの4つの分割に分けられる。まず、Flickr30K Entitiesの学習分割から、31個の未見語を選択し、これらの単語を含むすべてのテキスト-画像ペアを除外する。さらに、ホールドアウトされたテキスト-画像ペアは、未見語の学習セットとテストセットに分けられる。事前学習セットには、頻出単語(例:"man")が見単語のテスト結果を支配するのを防ぐため、Flickr30K Entitiesのテストスプリットから60個の見単語を選び、各単語について同数のテストインスタンスをサンプリングした。データセットの詳細と統計は付録Aを参照されたい。

dard hit-rate-at-$k$ (HR@$k$) measure and perplexity (Salazar et al., 2020; Jin et al., 2020). In masked language modeling, the log perplexity of a word $w$ is defined as the log pseudo-perplexity:

$$\log \text{PPL}(w) = -\log P(w|x_{\text{img}}, x_{\text{cap}}) \quad (1)$$

In object detection evaluation, especially for phrase grounding where multiple referents are possible (Kamath et al., 2021), Any-Protocol and All-Protocol are commonly adopted. Assuming $n$ ground truth bounding boxes $B = \{b_1, b_2, \cdots, b_n\}$ and $m$ predicted bounding boxes $\widetilde{B} = \{\widetilde{b_1}, \widetilde{b_2}, \cdots, \widetilde{b_m}\}$, the intersection-over-union (IoU) in both protocols is defined as:

$$\text{IoU}_{\text{any}} = \frac{1}{n} \sum_{i \in \{1,2,\cdots,n\}} \max_{j \in \{1,2,\cdots,m\}} \text{IoU}(b_i, \widetilde{b_j}) \quad (2)$$

$$\text{IoU}_{\text{all}} = \text{IoU}(\cup B, \cup \widetilde{B}) \quad (3)$$

However, these metrics only capture unimodal performance without concerning the correctness of cross-modal mapping. We design two new metrics to combine language and vision performance:

- **Grounded hit-rate** (G-HR@$k$), the proportion of tests with the masked word appearing in the top-$k$ candidates and a localization IoU over 0.5.
- **Grounded perplexity** (G-PPL) as follows:

$$\log \text{G-PPL}(w) = \begin{cases} \infty & \text{if IoU} = 0 \\ \log \text{PPL}(w) - \log \text{IoU} & \text{else} \end{cases} \quad (4)$$

### 2.3 Few-Shot Learning of New Words

Although there are grounding datasets available, *i.e.*, image-text pairs with word-object mapping annotation (Plummer et al., 2015), it is impractical to obtain such fine-grained annotation on a large scale and to cover the whole vocabulary space $\mathcal{V}$. We therefore explore grounded new word learning as a few-shot learning problem, especially under the setting of incremental class learning (Mandziuk and Shastri, 1999; Kemker et al., 2018). An intuitive illustration of the few-shot new word learning framework is provided in Figure 3. Under this framework, a computational model is developed in two stages. During the pre-training stage, the model receives image-caption pairs, with fine-grained word-object annotation for a set of base words $\mathcal{V}_{\text{seen}} \subseteq \mathcal{V}$. After pre-training, the model is provided with few samples of raw text-image pairs, each containing a set of unseen words $\mathcal{V}_{\text{unseen}} \subseteq \mathcal{V}$ that the model has to acquire.

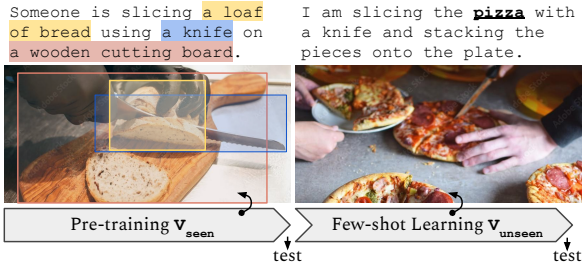



Figure 3: An illustration of the few-shot new word learning paradigm. The model first pre-trains on a grounding dataset with a set of base words ($\mathcal{V}_{\text{seen}}$), and then attempts to acquire a set of unseen words ($\mathcal{V}_{\text{unseen}}$) in a small number of raw text-image pairs. Tests are performed after each training session.

Tests are performed after each training stage. It's important to note that the unseen words may not be completely new, *e.g.*, the models may have encountered these words in its language encoder initialized with pre-trained language models. We consider them "unseen" because the model never sees these words paired with their referent, *i.e.*, the grounded meanings of the words are unknown.

### 2.4 Dataset Curation

We build our dataset based on the Flickr30K Entities dataset (Plummer et al., 2015), which contains image-text pairs with dense annotations between groundable phrases and bounding boxes of objects. The groundable phrases and regions are defined by the dataset, as chunks of text that refer to object bounding boxes. To construct word grounding instances, we use Stanza (Qi et al., 2020) to parse the caption, enumerate every word in the groundable phrase, and identify those with a POS tag of `NOUN` or `ADJ`. These groundable words are replaced by `MASK` one at a time and matched to their corresponding bounding boxes.

The dataset is divided into 4 splits: pre-training set, unseen words training set, seen words test set, and unseen words test set. We start by selecting 31 unseen words and holding out all text-image pairs containing these words from the training split of Flickr30K Entities. The hold-out text-image pairs are further divided into the training and test sets for unseen words. The remaining training split of Flickr30K Entities is used for the pre-training set. To prevent frequent words (*e.g.*, "`man`") from dominating the test results of the seen words, we choose 60 seen words and sample an equal number of test instances for each word from the test split of Flickr30K Entities. More details and statistics of the dataset are available in Appendix A.

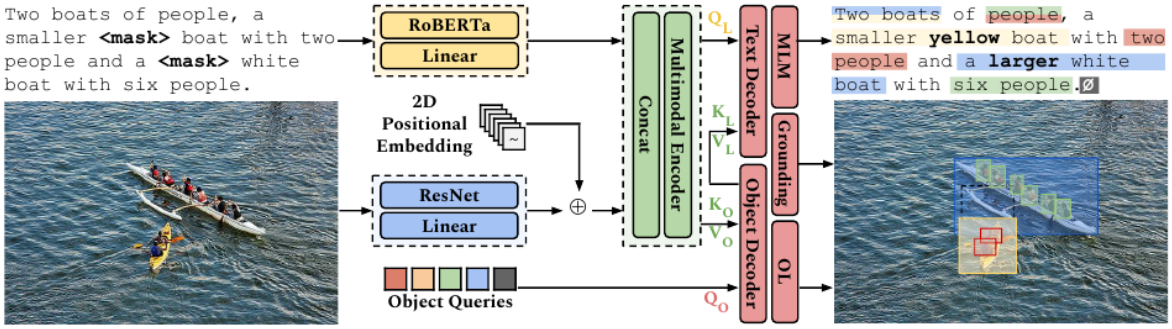



図4:W2Wアーキテクチャの概要。マスク言語モデリング(MLM)、物体位置特定(OL)、単語領域アライメントによる接地(WRA)の3つの目的で事前学習した視覚的根拠言語モデル。

## 3 Computational Models

### 3.1 The World-to-Words (W2W) Model

人間は、最小限の情報で新しい単語を学習する能力である高速なマッピングを実証している(Carey and Bartlett, 1978; Carey, 1978; Golinkoff et al., 2000)。我々は、視覚的な接地が新しい単語のブートストラップにどのように役立つかに動機づけられ、まず事前学習時に接地する能力を獲得し、接地した監視ができなくなったときにこの固有の能力を未知の単語の学習に移行させる計算フレームワークを提案する。我々は、図4に示すようなエンドツーエンドの設計を持つ新しい視覚的根拠に基づく言語モデルであるWorld-to-Words (W2W)を紹介する。

モデルアーキテクチャ。デュアルストリーム視覚言語モデルと同様に、W2Wはテキスト入力を事前に訓練された言語モデル(Liu et al., 2019)でエンコードし、画像入力を2D位置エンコーディングを追加した畳み込みバックボーン(He et al., 2016)でエンコードする。テキストと画像の表現は、結合された意味空間に線形に投影され、連結される。そして、マルチモーダル表現は自己注意層を持つクロスエンコーダに転送される。最終層のクロスエンコードされた表現は、学習可能なオブジェクトクエリのセットとともに、オブジェクトデコーダに送られる。オブジェクトデコーダは、各入力オブジェクトクエリに対してオブジェクト埋め込みを生成し、これは提案オブジェクトの表現とみなすことができる。オブジェクト表現はさらにテキストデコーダに転送され、言語モデリングが知覚されたオブジェクトに明示的に注目することを可能にする。前学習の目的、特に以下の段落でモデルがどのように根拠を獲得するかについて説明する。その他の詳細は付録Bに記載されている。

マスク言語モデリング(Masked Language Modeling: MLM)。本質的なタスクとして、既存の事前学習済み視覚言語モデルの大部分に従い、2層MLPによるマスク言語モデリングを実行する。入力テキスト中の単語はランダムにマスクされ、モデルは破損した文と画像を条件としてマスクされた単語を予測する。接地可能なフレーズ中の単語は0.4の確率で、接地不可能な領域中の単語は0.1の確率でマスクされる。

オブジェクトローカライゼーション(OL)。各オブジェクト表現は、共有の3層MLPによってデコードされ、バウンディングボックスが生成される。我々は、事前検出変換器(DETR)(Carionら、2020;Kamathら、2021)に従い、ハンガリー損失(Kuhn、1955)で提案ボックスとグランドトゥルースボックスの間の二分割マッチングを実行する。予測されたボックスは、一般化交差過剰結合(GIoU)損失(Rezatofighi et al.、2019)とL1損失を用いてグランドトゥルースに向かって最適化される。

接地。接地の概念は、単語とオブジェクトの間のきめ細かいクロスモーダルマッピングを可能にする単語領域アライメント(WRA)による接地型事前学習によって実現される。これは位置合わせと意味的アライメントの2つのレベルから構成される。位置合わせでは、モデルは各オブジェクト表現を文中の単語にマッピングすることを学習し、それはおそらくMASKまたは追加のオブジェクトなしラベル*(Yu and Siskind, 2013; Kamath et al.)
我々は、クロスエントロピー損失を用いてトークン位置の分布を予測するために、完全連結層を使用する。意味的アライメントでは、モデルは単語表現を接地先のオブジェクト表現に近づけ、関係のないペアを遠くに押し出すことを学習する。オブジェクトとテキストデコーダの最終層に対して対照的な損失を用いる。

3.2 ベースライン グランドレスベースライン。W2Wを同じ条件で事前学習し、損失関数に基底の目的を削除することで、基底能力のないベースラインを開発する。この基底なしモデルをW2W w/o Gと呼ぶ。

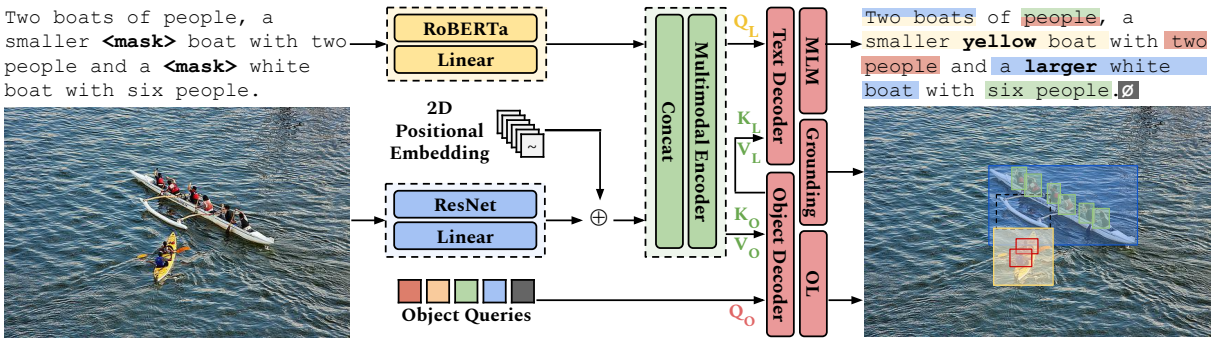



Figure 4: An overview of the W2W architecture, a visually grounded language model pre-trained with three objectives: masked language modeling (MLM), object localization (OL), and grounding through word-region alignment (WRA).

## 3 Computational Models

### 3.1 The World-to-Words (W2W) Model

Humans demonstrate fast mapping, the ability to learn new words with only minimal information (Carey and Bartlett, 1978; Carey, 1978; Golinkoff et al., 2000). Motivated by how visual grounding helps humans in bootstrapping new words, we propose a computational framework that first acquires the ability to ground during pre-training, and then transfers this intrinsic ability to learn unseen words when grounded supervision is no longer available. We introduce World-to-Words (W2W), a novel visually-grounded language model with an end-to-end design as illustrated in Figure 4.

**Model Architecture.** Similarly to dual-stream vision-language models, W2W encodes the textual input with a pre-trained language model (Liu et al., 2019), and encodes image input with convolutional backbone (He et al., 2016) with 2D positional encoding added. The text and image representations are linearly projected onto a joint semantic space and concatenated. The multimodal representation is then forwarded into a cross-encoder with self-attention layers. The cross-encoded representations in the final layer are sent into an object decoder, together with a set of learnable object queries. The object decoder produces an object embedding for each input object query, which can be considered as a representation of the proposed object. The object representations are further forwarded to the text decoder, which allows language modeling to explicitly attend to the perceived objects. We discuss the pre-training objectives, especially how the model acquires grounding in the following paragraphs. Other details are available in Appendix B.

**Masked Language Modeling (MLM).** As an intrinsic task, we follow the majority of existing pre-trained vision-language models to perform masked language modeling with a two-layer MLP. Words in input text are randomly masked out, and the model predicts the masked words conditioned on the corrupted sentence and image. Words in groundable phrases are masked with a probability of 0.4 and those in non-groundable regions are masked with a lower probability of 0.1.

**Object Localization (OL).** Each object representation will be decoded by a shared three-layer MLP to produce a bounding box. We follow prior detection transformers (DETR) (Carion et al., 2020; Kamath et al., 2021) to perform bipartite matching between proposed boxes and ground truth boxes with a Hungarian loss (Kuhn, 1955). The predicted boxes are optimized towards ground truth using the generalized intersection-over-union (GIoU) loss (Rezatofighi et al., 2019) and the L1 loss.

**Grounding.** The notion of *Grounding* is realized by grounded pre-training through word-region alignment (WRA) which enables fine-grained cross-modal mapping between words and objects. It consists of two levels of alignment: *positional alignment* and *semantic alignment*. In positional alignment, the model learns to map each object representation to words in the sentence, which could possibly be a MASK or an additional no-object label ∅ (Yu and Siskind, 2013; Kamath et al., 2021). We use a fully-connected layer to predict the distribution over token positions with cross-entropy loss. In semantic alignment, the model learns to bring word representations closer to the object representations that they ground to, and push the unrelated pairs farther. We use a contrastive loss over the final layers of the object and text decoders.

### 3.2 Baselines

**Groundless Baseline.** A baseline with no grounding ability is developed by pre-training W2W in the same condition but removing the grounding

| Models | Seen ($\|\mathcal{V}_{\text{seen}}\| = 60$) | | | | | | Unseen ($\|\mathcal{V}_{\text{unseen}}\| = 31$) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | G-HR@1 (↑) | log G-PPL (↓) | HR@1 (↑) | log PPL (↓) | Acc (↑) | IoU (↑) | G-HR@1 (↑) | log G-PPL (↓) | HR@1 (↑) | log PPL (↓) | Acc (↑) | IoU (↑) |
| RoBERTa | - | - | 38.0 | 2.75 | - | - | - | - | 23.1 | 4.96 | - | - |
| RoBERTa (FT) | - | - | 47.9 | 1.99 | - | - | - | - | 24.3 | 4.38 | - | - |
| ViLT | - | - | 64.7 | **1.27** | - | - | - | - | 32.7 | 3.68 | - | - |
| MDETR | - | - | - | - | 27.8 / 27.0 | 25.3 / 28.0 | - | - | - | - | 26.3 / 20.2 | 23.9 / 21.7 |
| ViLT+MDETR | 19.8 / 19.3 | 2.53 / 2.43 | 64.7 | **1.27** | 31.1 / 30.4 | 28.5 / 31.2 | 8.6 / 8.1 | 5.07 / 5.12 | 32.7 | 3.68 | 27.3 / 23.3 | 25.0 / 23.8 |
| VisualBERT (FT) | 28.5 / - | 2.96 / - | 42.3 | 2.33 | **68.1** / - | 53.3 / - | 10.2 / - | 5.60 / - | 20.7 | 4.81 | 50.6 / - | 45.2 / - |
| W2W$_{\text{w/o G}}$ (FT) | 28.9 / 27.8 | 2.33 / 2.38 | 63.9 | 1.41 | 44.0 / 43.0 | 40.0 / 38.2 | 1.1 / 1.1 | 11.89 / 12.04 | 3.7 | 10.87 | 38.7 / 31.9 | 36.2 / 31.0 |
| W2W | **47.0 / 46.3** | **1.79 / 1.81** | **66.9** | **1.26** | **66.8 / 66.3** | **58.8 / 57.6** | 2.3 / 2.3 | 11.58 / 11.74 | 4.2 | 11.01 | **61.3 / 53.1** | **56.3 / 48.0** |

表1:事前学習直後に得られた、見たことのある単語と見たことのない単語のテスト結果。微調整(FT)と明記されていない限り、全ての結果は微調整を行わないモデルの性能を反映している。AllとAny-protocolの評価は(All/Any)ペアとして表中に記載されている。事前学習済みの凍結物体検出器によるモデルについては、All-Protocolでの評価のみ可能である。なお、未見語はW2Wモデルでは未見語であり、事前学習済みのベースラインは開発中に全て遭遇している。参考までに結果を報告する。

典型的な事前学習済みVLM、例えばVisualBERT (Li et al., 2019)と同様に、W2W w/o Gは、明示的なクロスモーダル参照基底を用いず、オブジェクトの特徴に基づく言語モデリングを実行する。GOVAタスクに対して、収束するまで接地目的で事前学習データセット上でモデルを微調整することで、W2W w/o Gを適用する。

事前学習済みのベースライン 事前学習済みVLMの大部分において、事前学習時に未見の単語が既知である。また、本研究の主な目的は、グラウンディングと接地語取得のブートストラップを理解することである。これは、事前学習フレームワークのすべてのバリエーションをスケールアップしたり、再トレーニングしたりすることを目的としているわけではありません。そこで、参考と分析のためだけに、スケールが同等または適度に大きい事前学習済みVLMと我々のモデルを比較します。表1に示すように、フレーズグラウンディングの代表的なベースラインを選択した。

- "Detect-and-Recognize "ベースライン。このフレームワークのモデルは、事前に学習された凍結物体検出器に依存し、その後、提案された物体から単語を予測することを学習する。このタイプには、微調整されたVisualBERT (Li et al., 2019) を選択する。- "Produce-and-Localize "ベースライン。このフレームワークの下でのモデルは、欠落した単語を予測するために事前に訓練された視覚言語モデルに依存し、その後、参照表現理解を実行し、オブジェクトを提案する。ViLT (Kim et al., 2021) と MDETR (Kamath et al., 2021) を組み合わせることで、ビジョン条件付き言語モデリングとフレーズグラウンディングにおいて個別に競争力のある性能を発揮する。

## 4 Empirical Findings

### 4.1 Grounded Pre-training

本節の結果は、事前学習直後のテストから得られたものである。

| Models | # Param | # Imgs | # Caps | Objectives |
|---|---|---|---|---|
| RoBERTa | 120M | - | - | MLM |
| VisualBERT | 180M | 200K | 567K | MLM, ITM |
| ViLT | 110M | 4.0M | 10M | WRA*, MLM, ITM |
| MDETR | 200M | 200K | 1.3M | WRA, OL |
| W2W | 200M | 30K | 150K | WRA, MLM, OL |
| W2W$_{\text{w/o G}}$ | 200M | 30K | 150K | MLM, OL |

*WRAはViLTの単語パッチアライメントとして定式化されているため、大きな修正なしに物体位置特定を行うことができない。

表 2: 比較と参考文献のベースライン。ITM は Image Text Matching の略で、他の略語はすべてセクション 2 に従う。

見たことのある単語に対する事前学習結果 事前学習ステージの主な結果を表1にまとめた。W2Wは、Top-1 Grounded Hit-Rate (G-HR@1) とGrounded Perplexity (G-PPL) の両方の根拠となる指標において、W2Wの強力な性能を直接観察することができる。W2Wは、表2に示すように、より多くのデータと計算機で事前学習したシステムにおいても、GroundlessベースラインW2W w/o Gと事前学習済みベースラインを大幅に上回る性能を発揮する。W2Wは、欠落した単語の正しい予測と、対応するバウンディングボックスの位置を生成するが、ベースラインが両者を達成することは困難であることが判明した。Detect-and-Recognize」ベースライン(VisualBERT)では、凍結物体検出器によって強化された同等の物体定位性能が観察された。しかし、言語モデリング能力(HR@1やPPLで示されるように、RoBERTaの微調整よりも弱い)に悩まされている。Produce-and-Localize」ベースライン(ViLT+MDETR)では、ViLTの規模に起因する強い言語モデリング性能が観察された。しかし、正しい単語の接地は、ローカライズ性能の低さからわかるように、依然として困難である。これらの結果は、GOVAタスクが困難であり、W2Wが事前学習時に接地語の意味を学習することに競争力があることを示している。

Grounded Objectivesによるブートストラップ。さらに、事前学習効率におけるグラウンデッド・ターゲットの役割を過小評価するためのクロスタイム分析も行っている。

| Models | Seen ($\lvert\mathcal{V}_{\text{seen}}\rvert = 60$) | | | | | | Unseen ($\lvert\mathcal{V}_{\text{unseen}}\rvert = 31$) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | G-HR@1 (↑) | log G-PPL (↓) | HR@1 (↑) | log PPL (↓) | Acc (↑) | IoU (↑) | G-HR@1 (↑) | log G-PPL (↓) | HR@1 (↑) | log PPL (↓) | Acc (↑) | IoU (↑) |
| RoBERTa | - | - | 38.0 | 2.75 | - | - | - | - | 23.1 | 4.96 | - | - |
| RoBERTa (FT) | - | - | 47.9 | 1.99 | - | - | - | - | 24.3 | 4.38 | - | - |
| ViLT | - | - | 64.7 | **1.27** | - | - | - | - | 32.7 | 3.68 | - | - |
| MDETR | - | - | - | - | 27.8 / 27.0 | 25.3 / 28.0 | - | - | - | - | 26.3 / 20.2 | 23.9 / 21.7 |
| ViLT+MDETR | 19.8 / 19.3 | 2.53 / 2.43 | 64.7 | **1.27** | 31.1 / 30.4 | 28.5 / 31.2 | 8.6 / 8.1 | 5.07 / 5.12 | 32.7 | 3.68 | 27.3 / 23.3 | 25.0 / 23.8 |
| VisualBERT (FT) | 28.5 / - | 2.96 / - | 42.3 | 2.33 | **68.1** / - | 53.3 / - | 10.2 / - | 5.60 / - | 20.7 | 4.81 | 50.6 / - | 45.2 / - |
| W2W$_{\text{w/o G}}$ (FT) | 28.9 / 27.8 | 2.33 / 2.38 | 63.9 | 1.41 | 44.0 / 43.0 | 40.0 / 38.2 | 1.1 / 1.1 | 11.89 / 12.04 | 3.7 | 10.87 | 38.7 / 31.9 | 36.2 / 31.0 |
| W2W | **47.0 / 46.3** | **1.79 / 1.81** | **66.9** | **1.26** | **66.8 / 66.3** | **58.8 / 57.6** | 2.3 / 2.3 | 11.58 / 11.74 | 4.2 | 11.01 | **61.3 / 53.1** | **56.3 / 48.0** |

Table 1: Test results on the seen and unseen words, obtained immediately after pre-training. Unless noted explicitly as fine-tuned (FT), all results reflect the performance of models without fine-tuning. Evaluations under both All and Any-protocols are provided in the table as (All/Any) pairs. For models depending on a frozen pre-trained object detector, we can only provide evaluation under All-Protocol. We note that the unseen words are only unseen to W2W models, as pre-trained baselines have encountered them all during development. We report the results for reference.

objectives in the loss function. We refer to this groundless model as W2W$_{\text{w/o G}}$. Like a typical pre-trained VLM, *e.g.*, VisualBERT (Li et al., 2019), W2W$_{\text{w/o G}}$ performs language modeling based on the object features, without explicit cross-modal referential grounding. We apply W2W$_{\text{w/o G}}$ on GOVA task by fine-tuning the model on the pre-training dataset with grounding objective until convergence.

**Pre-trained Baselines.** For the majority of the pre-trained VLMs, the unseen words are known during pre-training. Also, the primary focus of this work is to understand grounding and bootstrapping in grounded word acquisition. It's not our goal to scale up or re-train all variants of pre-training frameworks. Therefore, we compare our model to the pre-trained VLMs with equal or reasonably larger scales for only reference and analysis purposes. We choose representative baselines in phrase grounding, as presented in Table 1:

- "Detect-and-Recognize" Baseline: Models under this framework rely on a pre-trained frozen object detector, and then learn to predict words from proposed objects. We choose the fine-tuned VisualBERT (Li et al., 2019) for this type.
- "Produce-and-Localize" Baseline: Models under this framework rely on a pre-trained vision-language model to predict the missing word, and then perform referring expression comprehension and propose objects. We combine ViLT (Kim et al., 2021) and MDETR (Kamath et al., 2021) for their competitive performance in vision-conditioned language modeling and phrase grounding individually.

## 4 Empirical Findings

### 4.1 Grounded Pre-training

The results of this section are obtained from the test immediately following pre-training.

| Models | # Param | # Imgs | # Caps | Objectives |
|---|---|---|---|---|
| RoBERTa | 120M | - | - | MLM |
| VisualBERT | 180M | 200K | 567K | MLM, ITM |
| ViLT | 110M | 4.0M | 10M | WRA*, MLM, ITM |
| MDETR | 200M | 200K | 1.3M | WRA, OL |
| W2W | 200M | 30K | 150K | WRA, MLM, OL |
| W2W$_{\text{w/o G}}$ | 200M | 30K | 150K | MLM, OL |

*WRA is formulated as word-patch alignment in ViLT, thus it cannot perform object localization without major modifications.

Table 2: The baselines for comparisons and references. ITM stands for Image Text Matching, and all the other abbreviations follow Section 2.

**Pre-training Results on Seen Words** The main results for the pre-training stage are summarized in Table 1. Our direct observation is the strong performance of W2W in terms of both grounded metrics, Top-1 Grounded Hit-Rate (G-HR@1) and Grounded Perplexity (G-PPL). W2W significantly outperforms the groundless baseline W2W$_{\text{w/o G}}$ and pre-trained baselines, even for systems pre-trained with a significantly larger amount of data and computing, as shown in Table 2. While W2W produces correct predictions of the missing words as well as the locations of the corresponding bounding boxes, it turns out to be challenging for baselines to achieve them both. For "Detect-and-Recognize" baseline (VisualBERT), we observe a comparable object localization performance empowered by the frozen object detector. However, it suffers from a poor language modeling ability (as demonstrated by HR@1 and PPL, weaker than a fine-tuned RoBERTa). For the "Produce-and-Localize" baseline (ViLT+MDETR), we observe a strong language modeling performance due to the scale of ViLT. Yet, correct word grounding remains difficult, as can be seen from the poor localization performance. These results demonstrate that the GOVA task is challenging, and W2W is competitive in learning grounded word meanings during pre-training.

**Bootstrapping through Grounded Objectives.** We further provide a cross-time analysis to under-

異なる学習ステップの結果を表3に示す。表から、W2Wは言語モデリング、オブジェクトローカライゼーション、および、グラウンデッドパープレキシティの下での共同学習において、そのアースなしの両バリアントを上回ることが観察される。さらに顕著なのは、W2Wは、グラウンデッド目的語なしで学習したモデル(すなわち、WRA目的語)と比較して、10倍少ない学習データでより良い性能を達成することである。これらの結果は、効率的な接地語学習において、明示的な単語とオブジェクトのアライメントが重要な役割を果たすことを裏付けています。これは、接地目的が視覚と言語の意味空間を整合させようとすることで説明でき、視覚的条件付き言語モデリングと言語条件付き物体位置特定の両方に理想的に利益をもたらす。また、事前学習後のクロスモーダルプロービングと微調整により、単語表現と物体表現のマッピングを構築することは可能であるが、これらの方法は、効率と性能の点で、接地目的語を持つシステムと比較することはできない

| # Steps | Metrics | W2W | $W2W_{w/o\ G}$ (FT) |
|---|---|---|---|
| 10k | IoU (↑) | **46.7 / 46.2** | 36.9 / 35.3 |
| | log PPL (↓) | **1.46** | 1.53 |
| | log G-PPL (↓) | **2.22 / 2.23** | 2.52 / 2.57 |
| 50k | IoU (↑) | **58.1 / 57.1** | 39.6 / 38.8 |
| | log PPL (↓) | **1.26** | 1.44 |
| | log G-PPL (↓) | **1.80 / 1.82** | 2.34 / 2.38 |
| 100k | IoU (↑) | **58.7 / 57.6** | 40.0 / 38.2 |
| | log PPL (↓) | **1.26** | 1.41 |
| | log G-PPL (↓) | **1.79 / 1.81** | 2.34 / 2.38 |

表3:異なる学習ステップにおけるW2Wとその非接地版の比較。$W2W_{w/oG}$はfine-tuningで評価。AnyとAll-protocolsは(All/Any)のペアとして表中に記載されている。

未見語に対する事前学習結果。単語にとらわれない接地事前学習モデルの重要な発見は、MASKの背後にある未見語をローカライズする際の驚くべき性能である。表1に示すように、W2Wは未見語に対して56.3%という高いAnyIoUと61.3%というAny-localization精度を達成し、これは見た集合での性能に非常に近く、これらの単語を見たベースラインを上回っています。さらに、予想通り、これらの単語は事前学習中に保留されるため、W2Wはこれらの未見単語を正しくマスクすることができず、見た単語の対数パープレキシティが1.26と66.9と比較して11.01と高く、HRが4.2と低くなっています。図5は、このような単語に依存しない接地方法の一例である。


図5:W2Wでは「象牙」という単語は見られないが、MASKで参照される画像中の物体を局在化させることができるモデルである。

このように言語モデリングと未見語に対する参照元の位置決めに性能差があることは、W2Wが、事前学習中に単語そのものを見ることがなくても、言語文脈と視覚文脈の両方を通じて、あるレベルの単語にとらわれない根拠付け、すなわち、最も可能性の高い単語の参照元を特定することを発達させたことを示唆している。同様の状況は、図1で前述したように、人間の言語学習者が新しい単語の根拠となる意味を推論する際にも直面する。我々の実験では、根拠のある事前学習により、視覚言語システムが単語にとらわれない根拠付け能力を獲得することが可能であり、新しい単語を学習する際に人間のような高速マッピングを可能にする機会を開くことができることを実証している。

### 4.2 Few-Shot New Words Acquisition

本節では、W2Wが少数の生画像-テキストペアのサンプルから未見語を取得するようにタスク付けを行い、バウンディングボックスや単語-オブジェクトマッピングのアノテーションを行わない。我々は、このモデルの単語にとらわれない根拠を示したので、大量のデータと根拠のある監視が利用できなくなったときに、この能力を移して未見語の学習を促進できるかどうかを探ろうとするものである。具体的には、マスク言語モデリング(MLM)のみを学習目標として、事前に学習したW2Wに対して数発の学習を行う。ハイパーパラメータの詳細は付録B.2を参照されたい。

インクリメンタル学習による新しい単語の学習。まず、多クラス漸増学習の設定について検討する。この設定では、事前学習したモデルに、数回の学習セッションで未見の31個の単語を獲得するタスクが課される。この実験は、事前学習直後にサンプルサイズ8、16、24、32で繰り返される。図6に示すように、1単語あたり8サンプルでも、W2Wは未見語の接地語のパープレキシティを大幅に低下させることができ、ほとんどが破局的忘却なしに見た語の接地語のパープレキシティを維持することができました。また、W2WはW2Wと比較して、未見語に対してより良い獲得性能を示す。

stand the role of grounded objectives in pre-training efficiency. The results of different training steps are provided in Table 3. From the table, we observe that W2W outperforms both of its groundless variants in language modeling, object localization, and jointly under the grounded perplexity. What's even more striking is that W2W achieves better performance with *10 times less training data* compared to the model trained without the grounding objective (*i.e.*, the WRA objective). These results confirm the crucial role of explicit word-object alignment in efficient grounded word learning. This can be explained by that the grounded objectives attempt to align the vision and language semantic spaces, which ideally benefit both visually conditioned language modeling and language-conditioned object localization. Although it is possible to build a mapping between word and object representations through cross-modal probing and fine-tuning after pre-training, these methods are not comparable to systems with grounded objectives in terms of efficiency and performance.

| # Steps | Metrics | W2W | $\text{W2W}_{\text{w/o G}}$ (FT) |
|---|---|---|---|
| 10k | IoU (↑) | **46.7 / 46.2** | 36.9 / 35.3 |
| | log PPL (↓) | **1.46** | 1.53 |
| | log G-PPL (↓) | **2.22 / 2.23** | 2.52 / 2.57 |
| 50k | IoU (↑) | **58.1 / 57.1** | 39.6 / 38.8 |
| | log PPL (↓) | **1.26** | 1.44 |
| | log G-PPL (↓) | **1.80 / 1.82** | 2.34 / 2.38 |
| 100k | IoU (↑) | **58.7 / 57.6** | 40.0 / 38.2 |
| | log PPL (↓) | **1.26** | 1.41 |
| | log G-PPL (↓) | **1.79 / 1.81** | 2.34 / 2.38 |

Table 3: Comparison of W2W and its non-grounding version at different training steps. $\text{W2W}_{\text{w/o G}}$ is evaluated using fine-tuning. Both Any and All-protocols are provided in the table as (All/Any) pairs.

**Pre-training Results on Unseen Words: Word-Agnostic Grounding** One important finding of the pre-trained model is the surprising performance in localizing the unseen words behind the MASKs. As shown in Table 1, W2W achieves a high Any-IoU of 56.3% and Any-localization accuracy of 61.3% for the unseen words, which are very close to its performance on the seen set and surpass baselines that have seen these words. Moreover, as anticipated, since these words are held out during pre-training, W2W fails to correctly unmask these unseen words, leading to a high log perplexity of 11.01 and low HR of 4.2, compared to that of 1.26 and 66.9 on the seen words. Figure 5 shows an example of such word-agnostic grounding.

Three men seated on a **<MASK>** in a small village.

- W2W: animal
- Ground Truth: elephant

Figure 5: Although the word "elephant" is unseen to W2W, the model is still able to localize the object in the image referred to by the MASK.

This performance disparity in language modeling and referent localization on unseen words suggests that W2W has developed a certain level of word-agnostic grounding, *i.e.*, to locate the most likely referent of a word through both the linguistic context and the visual context, even if the word itself is never seen during pre-training. A similar situation is faced by human language learners when inferring the grounded meaning of a new word, as we described earlier in Figure 1. Our experiment demonstrates that, through grounded pre-training, it is possible for a vision-language system to acquire word-agnostic grounding ability, which opens up the opportunity to enable human-like fast mapping when learning new words.

### 4.2 Few-Shot New Words Acquisition

In this section, we task W2W to acquire unseen words from a few samples of raw image-text pairs, without any bounding boxes or word-object mappings annotation. As we have demonstrated the model's word-agnostic grounding, we seek to explore if this ability can be transferred to facilitate learning unseen words when a large amount of data and grounded supervision are no longer available. Specifically, we perform few-shot learning on the pre-trained W2W with only masked language modeling (MLM) as the learning objective. More hyperparameter details are available in Appendix B.2.

**Learning New Words through Incremental Learning.** We first explore the multi-class incremental learning setting, in which the pre-trained model is tasked to acquire the 31 unseen words from a few-shot learning session. The experiment is repeated with sample sizes of 8, 16, 24, and 32 immediately after pre-training. As shown in Figure 6, even with as few as 8 samples per word, W2W can significantly bring down the grounded perplexity of unseen words, while mostly maintaining the grounded perplexity of the seen words without catastrophic forgetting. Compared to W2W without the grounding objective, the full W2W demonstrates better acquisition performance for unseen words.

これらの数個のショット例は、明示的な根拠付けのないテキスト/画像ペアであることに注意することが重要である。我々のW2Wは、新しい単語の根拠となる意味を(例えば、8つの例でのみ)素早く獲得することができ、見たことのある単語に近い性能を持つ。

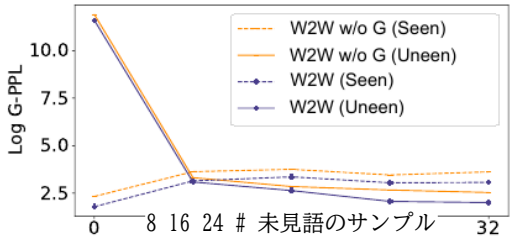



図6:多クラス漸増学習における、サンプルサイズ8から32までの未見語の対数G-PPL(All-Protocol)。

さらに、1クラス漸増学習設定による単語特異的な対照研究を行う。表4は、2つの未見語(ピザと円形)に対する結果を示している。完全な結果は付録Dに掲載されている。

| # Samples | log G-PPL (pizza) | | log G-PPL (circular) | |
|---|---|---|---|---|
| | W2W | $W2W_{w/o\ G}$ | W2W | $W2W_{w/o\ G}$ |
| 0 | 10.70 | 9.59 | 15.21 | 15.12 |
| 8 | **1.47** | 2.21 | **1.59** | 2.25 |
| 16 | **1.07** | 2.54 | **1.07** | 2.25 |
| 24 | **1.19** | 1.25 | **1.55** | 1.81 |
| 32 | **0.90** | 1.18 | **1.23** | 1.61 |

表4:1クラス漸増学習における未見語のlog G-PPL (All-Protocol)、各未見語のサンプルサイズは8から32の範囲。

### 4.3 モデル動作の予測因子

事前学習された言語モデルの性能や振る舞いを説明し/予測できる予測因子を特定することに関心が集まっている(Chang and Bergen, 2022)。この探索は、将来のモデル開発に貴重な洞察を与えるだけでなく、言語モデルが人間の言語習得パターンとどの程度整合しているかを評価するための認知的探究としても機能する。本節では、視覚言語モデルに関するこの性質に関する最初の研究成果を紹介する。具体的には、W2WモデルはRoBERTaエンコーダに依存しており、これは既に事前言語知識を備えていた可能性があることに注意する。また、視覚言語モデルと人間の言語習得の認知的整合性を評価するために、ランダムに初期化したRoBERTaエンコーダを用いてW2WとW2Wのw/o Gモデルを事前学習させた。

単語の様々な側面を包括的に捉えるために、我々は、固有の心理言語学的特性、訓練コーパス内の分布パターン、訓練画像内の視覚表現を包含する8つの異なる予測因子を慎重に選択した。我々は、MRCデータベース(Coltheart, 1981)から収集・正規化した3つの心理言語学的予測因子を選択した。- Familiarity:人々が言葉に慣れているか、触れているかの度合い、Concreteness:言葉が知覚できる物理的な参照元を持っているか、有形のオブジェクトや経験と関連しているか、 - Imageability:言葉が人々の心的イメージを引き出す度合い。

もう一つの3つの言語予測変数が検討されている。- Unigram perplexity; - RoBERTa perplexity: RoBERTaがキャプション上で微調整され、ユニモーダル言語モデル性能の上限となる; - # Co-occur phrases: キャプション内で共起するgroundable phrasesの平均数。

最後に、2つの知覚的予測因子を選択する。- Bbox size: 参照オブジェクトのバウンディングボックスによって占有される画像の平均割合。

各予測変数の統計的有意性を評価するために、異なるモデルのバリエーションについて尤度比検定による線形回帰を行った。Chang and Bergen (2022)と同様に、ターゲット予測変数を含む全体的な回帰と、ターゲット以外のすべての予測変数を含む回帰を比較する。さらに、相関の大きさと方向を捉えるために、ベータ重み(符号付き)を提示する。図7は、Log G-PPL、Log PPL、Any IoUに関する各予測変数の統計的有意性(負の対数p値で)を示すヒートマップである。有意でない検定は図から省略した。

言語予測器および知覚予測器との相関。その結果、groundedとungroundedのシナリオの両方において、unigramとRoBERTaの対数パープレキシティとモデルの対数パープレキシティの間に正の相関があることが明らかになった。このことは、視覚言語モデルは、ユニモーダルモデルと同様に、依然として分布統計量に大きく依存していることを示している。

It's important to note that these few shot examples are text/image pairs without explicit grounding annotation. Our W2W is able to quickly acquire grounded meanings of the new words (*e.g.*, only with 8 examples) with a performance close to that of seen words.

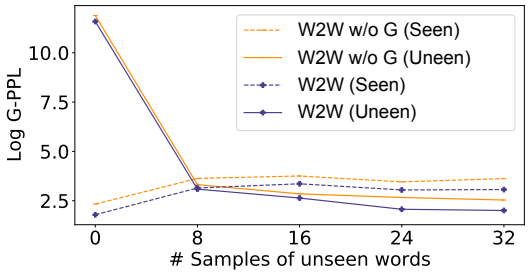



Figure 6: The log G-PPL (All-Protocol) of seen and unseen words in multi-class incremental learning, each unseen word with a sample size ranging from 8 to 32.

We further perform a word-specific controlled study with a one-class incremental learning setting. We present results on two unseen words (`pizza` and `circular`) in Table 4. The complete results are available in Appendix D.

| # Samples | log G-PPL (`pizza`) | | log G-PPL (`circular`) | |
|---|---|---|---|---|
| | W2W | $\text{W2W}_{\text{w/o G}}$ | W2W | $\text{W2W}_{\text{w/o G}}$ |
| 0 | 10.70 | 9.59 | 15.21 | 15.12 |
| 8 | **1.47** | 2.21 | **1.59** | 2.25 |
| 16 | **1.07** | 2.54 | **1.07** | 2.25 |
| 24 | **1.19** | 1.25 | **1.55** | 1.81 |
| 32 | **0.90** | 1.18 | **1.23** | 1.61 |

Table 4: The log G-PPL (All-Protocol) of unseen words in one-class incremental learning, each unseen word with a sample size ranging from 8 to 32.

### 4.3 Predictors of Model Behaviors

There has been an interest to identify predictors that can explain/anticipate the performance or behavior of pre-trained language models (Chang and Bergen, 2022). This exploration not only offers valuable insights for future model development, but also serves as a cognitive inquiry to evaluate the extent to which language models align with human language acquisition patterns. In this section, we present the first work of this nature on vision-language models. Specifically, we note that the W2W model relies on a RoBERTa encoder, which might have already been equipped with prior linguistic knowledge. To assess the cognitive alignment of vision-language models to human language acquisition, we additionally pre-trained the W2W and $\text{W2W}_{\text{w/o G}}$ models with a randomly initialized RoBERTa encoder.

To comprehensively capture various aspects of words, we carefully select eight distinct predictors that encompass intrinsic psycho-linguistic characteristics, distribution patterns within the training corpus, and visual representations within the training images. We select 3 **psycho-linguistic predictors**, each collected and normalized from the MRC Database (Coltheart, 1981):

- `Familiarity`, the degree of familiarity or exposure people have to words;
- `Concreteness`, the degree to which words have a perceptible physical referent or are associated with tangible objects or experiences;
- `Imageability`, the degree to which words elicit people's mental imagery.

Another 3 **linguistic predictors** are considered:

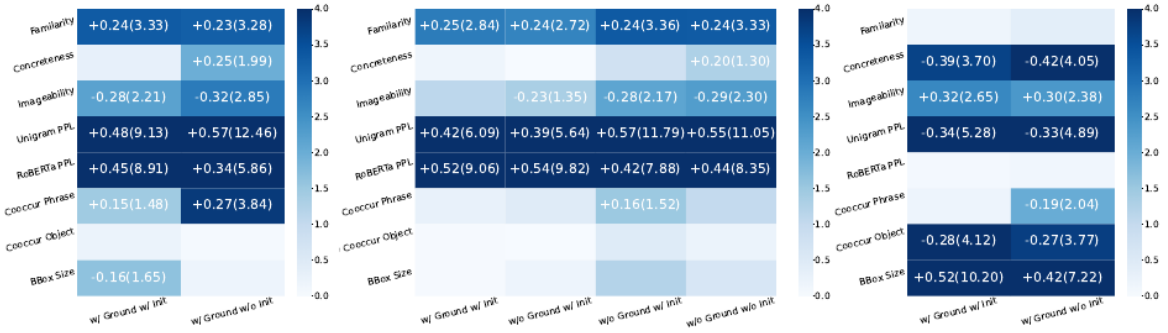
- `Unigram perplexity`;
- `RoBERTa perplexity`, where RoBERTa is fine-tuned on the captions to serve as the upper bound of unimodal language model performance;
- `# Co-occur phrases`, the average number of co-occurring groundable phrases in a caption.

We finally choose 2 **perceptual predictors**:

- `# Co-occur objects`, the average number of co-occurring objects in an image;
- `Bbox size`, the average proportion of an image occupied by the bounding boxes of the referents.

To assess the statistical significance of each predictor, we performed linear regressions with likelihood ratio tests on different variants of models. Similar to Chang and Bergen (2022), we compare the overall regression including the target predictor to a regression that included all predictors except the target. We additionally present the beta weights (with signs) to capture the magnitude and direction of the correlation. Figure 7 displays heatmaps indicating the statistical significance (in terms of negative logarithmic $p$-values) of each predictor concerning Log G-PPL, Log PPL, and Any IoU. Insignificant tests are omitted from the figure.

**Correlation with Linguistic and Perceptual Predictors.** Our findings revealed a positive correlation between the unigram and RoBERTa log perplexity and the models' log perplexity, both for grounded and ungrounded scenarios. This indicates that vision-language models still heavily rely on distributional statistics, similar to unimodal models. While the ungrounded perplexity showed little correlation with perceptual predictors, the Any




(a) Predictors for Log G-PPL. (b) Predictors for Log PPL. (c) Predictors for Any IoU.

図 7: Log G-PPL, Log PPL, Any IoU に対する各予測変数の統計的有意性のヒートマップ。括弧の外側にはベータウェイトとその符号が、括弧には負の対数 p 値が示されている。p > 0.05、すなわち{-} log(p) < 1.30 の有意でない検定は破棄される。

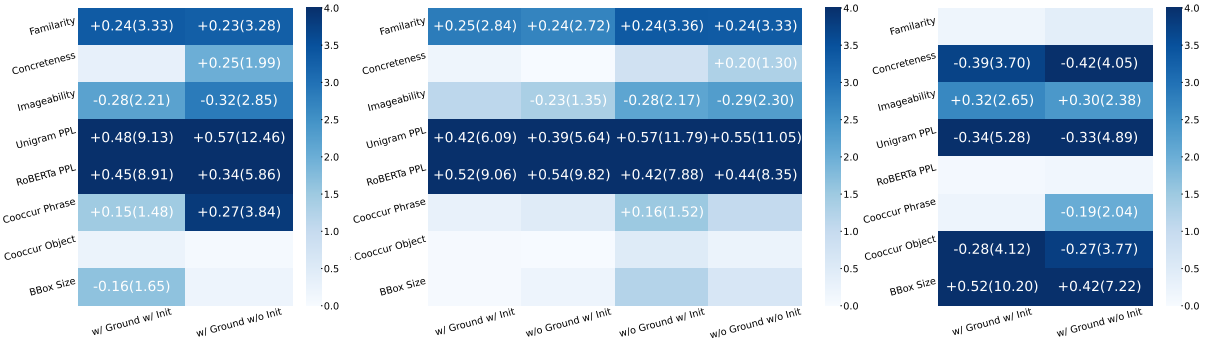
非接地性パープレキシティは知覚予測因子とほとんど相関がなかったが、Any IoUは共起するオブジェクトの数およびバウンディングボックスの平均サイズと有意な相関を示した。このことは、視覚的に顕著で知覚的に曖昧でない概念は、人間の学習者と一致し、ローカライズと獲得が容易であることを示唆している(Smith and Yu, 2008)。

心理言語学的予測因子との相関。直感に反して、人間の言葉の親しみやすさと機械の複雑さの間には正の相関があり、つまり、人間が言葉に慣れ親しんでいるほど、モデルはより困惑している。これは、人間の言語習得の理想的な認知的妥当性とは対照的である。この矛盾は、現在の視覚言語モデルが認知的妥当性を十分に達成できていない可能性を示唆している。これは、多くの概念(例えば、野生動物、楽器)がインターネット画像には豊富に現れるが日常生活には現れないという事実によって説明されるかもしれない。イメージング性という点では、人間の直感とよく一致し、Any IoUと正の相関を示し、当惑度とは負の相関を示した。しかし、具体性予測因子は意外にも逆の相関を示した。この不一致は、イメージング性と具体性の微妙な区別に起因している可能性がある。例えば、「帽子」は具体的な物体を指すため具体的であるが、その一般性から視覚的な多様性も有意であり(例えば、見た目が大きく異なる多くの種類の帽子)、習得が困難である。逆に、「青」は特定の具体的な対象に言及していないにもかかわらず、比較的安定した色を容易に呼び起こすので、よりイメージしやすくなります。帽子の意味を学ぶために、人間の言語学習者は、物体と物理的に相互作用することで利益を得ることができ、帽子は視覚的な外観に関係なく、頭を覆うためのアイテムであることを理解することができます。

このギャップを解決するために、将来的には、受動的な知覚ではなく、物理的な相互作用によって単語を獲得する言語学習エージェントを開発し、単語の意味をより包括的に理解することが必要となる可能性があります。

## 5 Related Work

視覚-言語マッピングマッピングマッピングは古典的な語彙習得問題において中心的な役割を果たす(Gleitman and Landau, 1994; Clark, 1995)。主に、単語を意味記号に接地し、特定の精神的バイアスを用いて子どもの単語習得をシミュレートする学習メカニズムを構築し、心理言語現象の計算論的説明を与えることに焦点を当てた(Siskind, 1996; Regier, 2005; Goodman et al, 2007; Fazly et al, 2010)。この線に沿った初期の取り組みは、オブジェクトカテゴリ(Roy and Pentland, 2002; Yu, 2005; Xu and Tenenbaum, 2007; Yu and Ballard, 2007; Yu and Siskind, 2013)やより複雑な視覚特徴(Qu and Chai, 2010; Mao et al, 2019, 2021; Pratt et al, 2020)から言語ラベルへの統計的または神経マッピングを学習することによって視覚的根拠を組み込んでいる。これらの研究は通常、語彙が限られた閉じた世界で行われ(Krahmer and van Deemter, 2019)、g., object retrieval (Guadarrama et al., 2014; Hu et al., 2016), referring expression understanding and grounding (Liu et al., 2014; Yu et al., 2016; Mao et al., 2016; Wu et al., 2020), and phrase grounding (Prummer et al., 2015) は参照表現を対応するオブジェクトにマッピングする。我々の設定は、接地を明示的な単語参照マッピング問題として位置づけるため、この境界線と密接に関連している。




(a) Predictors for Log G-PPL. (b) Predictors for Log PPL. (c) Predictors for Any IoU.

Figure 7: Heatmaps for statistical significance for each predictor towards the Log G-PPL, Log PPL, and Any IoU. The beta weights and their signs are presented outside of the parentheses, and the negative log $p$-values are presented in the parentheses. Insignificant tests with $p > 0.05$, *i.e.*, $-\log(p) < 1.30$, are discarded.

IoU demonstrated a significant correlation with the number of co-occurring objects and average sizes of bounding boxes. This suggests concepts that are visually salient and less perceptually ambiguous are easier to localize and acquire, consistent with human learners (Smith and Yu, 2008).

**Correlation with Psycho-linguistic Predictors.** Counter-intuitively, there was a positive alignment between the human perceived familiarity of words and the machine's perplexities, *i.e.*, the more familiar humans are with a word, the more perplexed models get. This contrasts with the ideal cognitive plausibility of language acquisition in humans. This discrepancy implies that current vision-language models may not fully achieve cognitive plausibility, which might be explained by the fact that many concepts (*e.g.*, wild animals, musical instruments) appear abundantly in internet images but not in daily lives. In terms of imageability, it aligned well with human intuition, exhibiting a positive correlation with Any IoU and a negative correlation with perplexities. However, the concreteness predictor surprisingly exhibited the opposite correlation. This discrepancy could be attributed to the nuanced distinction between imageability and concreteness. For instance, while "`hat`" is concrete because it refers to a tangible object, it also possesses visual diversity due to its generality (*e.g.*, many types of hats which look very differently), making it challenging to acquire. Conversely, "`blue`" is more imageable as it easily evokes a color, relatively stable, despite not referring to a specific tangible object. To learn the meaning of "`hat`," a human language learner may benefit from physically interacting with the object, and understand that the hat is an item to cover for the head, regardless of its visual appearance. To address this gap, a potential future direction could involve developing language learning agents that acquire words through physical interactions rather than passive perception, allowing for a more comprehensive understanding of word meanings.

## 5 Related Work

**Vision-Language Mapping** Mapping plays a central role in classic lexicon acquisition problem (Gleitman and Landau, 1994; Clark, 1995). Primarily, researchers focused on grounding words to their meaning symbols, building learning mechanisms using specific mental biases to simulate children's word acquisition, and giving computational accounts for psycholinguistic phenomena (Siskind, 1996; Regier, 2005; Goodman et al., 2007; Fazly et al., 2010). Early efforts along this line incorporate visual grounding either by learning a statistical or neural mapping from object categories (Roy and Pentland, 2002; Yu, 2005; Xu and Tenenbaum, 2007; Yu and Ballard, 2007; Yu and Siskind, 2013) and more complicated visual features (Qu and Chai, 2010; Mao et al., 2019, 2021; Pratt et al., 2020) to linguistic labels. These studies are usually in a closed world with limited vocabulary (Krahmer and van Deemter, 2019), and words are usually isolated from the natural context of use. More recently, multi-modal understanding tasks, *e.g.*, object retrieval (Guadarrama et al., 2014; Hu et al., 2016), referring expression comprehension and grounding (Liu et al., 2014; Yu et al., 2016; Mao et al., 2016; Wu et al., 2020), and phrase grounding (Plummer et al., 2015) map referring expressions to corresponding objects. Our setup is closely related to this line as we position *grounding* as an explicit word-referent mapping problem. The difference is that, our work goes beyond grounding

しかし、我々の研究は、AIエージェントが直面するより複雑だが現実的な課題である、高速マッピングによるオープンボキャブラリーの獲得を研究するために、接地を超えるものである。

視覚言語の事前学習 分布的な単語表現は言語モデリングによって獲得することができ、視覚データからの言語モデルの開発はコミュニティによって広く研究されている(Chrupała et al.2015; Lazaridou et al.2015; Li et al.2017; Surıs et al.2020 )。近年、視覚拡張言語モデリングで言語表現を豊かにする研究(Tan and Bansal, 2020; Lu et al., 2022; Wang et al., 2022)や、視覚言語事前学習(VLP)でマルチモーダル表現を学習する研究(Du et al., 2022a)が増えてきている。我々は特に、Word-Region Alignment (WRA)などの細かい根拠を持つVLPモデルに関心がある。これらのモデルは、パッチと単語をマッチングさせる最適輸送のような弱教師付きアライメントアルゴリズムで事前学習するか(Kimら、2021)、凍結検出器からの提案(Chenら、2020;Suら、2022)のいずれかである。, 2020)、あるいは言語条件付き検出器の事前学習により明示的な単語接地を行う(Kamath et al., 2021; Li et al., 2022; Zhong et al., 2022; Dou et al., 2022)。我々のモデルはこの線上に位置し、既存の物体検出器に依存するのではなく、事前学習時に言語モデリング、物体定位、接地を共同で行う。

視覚言語タスク 視覚言語システムを評価するために、多くの下流タスクが定式化されている。いくつかの関連する定式化は付録の表5にまとめられている。これらのダウンストリームタスクは、いくつかの視覚言語能力を実証しているが、これらのモデルが外部環境に対する単語の根拠ある意味を本当に捉えているかどうかについての洞察は限られている。我々のタスクデザインは、特に機械が単語を予測し、単語を知覚に向ける能力をターゲットとしている。我々の定式化に近いのは、継続的な学習設定におけるビジョンベースの言語モデリングタスク(Jin et al.、2020)である。我々の研究は主に2つの点で異なっている。まず、Jinら(2020)が提案したタスクは、視覚的文脈に基づいてマスクされたトークンのみを予測するため、参照の不確実性(すなわち、接地)はそのままにされる(例えば、図2において、「船」という単語の正しい予測は、正しい接地を保証しない)。また、本研究は主に構成性に着目しているが、未見語に遭遇した場合の数発の接地語学習を目指すオープンボキャブラリー物体検出 初期の研究では、新しい単語の高速マッピングをゼロショット物体分類問題として定式化し、既知の物体ラベルから未知の物体ラベルに一般化することを目的としている(Socherら、2013;Fromeら、2013;Elhoseinyら、2013;Lazaridouら、2014)。

after pre-training.

この設定は後にゼロショット物体検出(ZSD)と呼ばれるローカライゼーションタスクに拡張される(Bansal et al., 2018; Zhu et al., 2019, 2020; Rahman et al., 2020)。より最近では、オープンボキャブラリーオブジェクト検出(OVD)(Zareianら、2021;Guら、2022;Duら、2022b;Mindererら、2022)が、従来のゼロショット設定の非現実的な制約に対処するために、ZSDと弱教師付きオブジェクト検出(WSD)を組み合わせている。OVDは粗視化された画像とキャプションのペアが利用可能であることを前提とし、オブジェクトカテゴリの限られた細粒度のアノテーションから未見のものへの汎化を試みている。しかし、この研究では、単語をオブジェクトカテゴリとして位置づけ、言語的文脈(例えば、文)から単語を分離している。本研究では、人間が生成したキャプションの言語モデリングを行うためのモデルに挑戦する。

## 6 Conclusion and Future Work

言語とその参照語のつながりは、単語の根拠となる意味を捉えており、明示的な処理は、人間やAIエージェントの効率的なオープンワールド言語学習能力を強化するための鍵である。本研究では、オープンワールドの接地言語学習における接地と高速マッピングを検討するためのスケーラブルな定式化であるGrounded Open Vocabulary Acquisition (GOVA)を紹介する。我々は、視覚的根拠に基づく新しい言語モデルであるWorld-to-Words (W2W)を提案し、モデルが事前学習中に最初に接地能力を獲得し、その後、明示的な根拠付けの監視なしに新しい単語を迅速に学習するためにこの能力を適用するパラダイムを調査する。我々の実証的な発見は、神経単語の獲得における視覚的根拠の重要性を強調するものである。特に、事前に学習したW2Wは、数発学習により新規接地語を高速にマッピングするための基盤となることを見出した。また、視覚言語モデルの性能に影響を与える潜在的な予測因子を探索するために包括的な分析を行い、人間の言語学習パターンに関して一貫した行動と驚くべき行動の両方を明らかにする。これらの知見は、

to study open-vocabulary acquisition through fast mapping, a more complicated but realistic challenge faced by AI agents.

**Vision-Language Pre-training** Distributional word representations can be acquired through language modeling, and developing language models from visual data has been extensively studied by the community (Chrupała et al., 2015; Lazaridou et al., 2015; Li et al., 2017; Surıs et al., 2020). Recent years have seen increasing research to enrich language representations with visually-augmented language modeling (Tan and Bansal, 2020; Lu et al., 2022; Wang et al., 2022) and to learn multimodal representations with vision-language pre-training (VLP) (Du et al., 2022a). We are particularly interested in VLP models with fine-grained grounding objectives, *e.g.*, Word-Region Alignment (WRA). These models either pre-train with weakly supervised alignment algorithms like optimal transport that matches words with patches (Kim et al., 2021) or proposals from a frozen detector (Chen et al., 2020; Su et al., 2020), or perform explicit word grounding by pre-training a language-conditioned detector (Kamath et al., 2021; Li et al., 2022; Zhong et al., 2022; Dou et al., 2022). Our model falls along this line, which jointly performs language modeling, object localization, and grounding during pre-training, rather than relying upon a pre-existing object detector.

**Vision-Language Tasks** To evaluate vision-language systems, many downstream tasks have been formulated. Some related formulations are summarized in Table 5 in Appendix. While demonstrating some vision-language capabilities, these down-stream tasks provide limited insights into whether these models truly capture the grounded meaning of words with respect to the external environment. Our task design specifically targets the machine's ability to predict words and ground words to perception. More akin to our formulation is the vision-based language modeling task (Jin et al., 2020) in a continual learning setting. Our work differs mainly in two aspects. First, the task proposed by Jin et al. (2020) only predicts masked tokens based on the visual context, which leaves the referential uncertainty (i.e., grounding) unattended (*e.g.*, in Figure 2, correct prediction of the word "`boat`" does not guarantee correct grounding). Also, this work primarily focuses on compositionality, while we seek to address few-shot grounded word learning when unseen words are encountered after pre-training.

**Open-Vocabulary Object Detection** Early works formulate fast mapping of new words as a zero-shot object classification problem, which aims to generalize from known object labels to unknown ones (Socher et al., 2013; Frome et al., 2013; Elhoseiny et al., 2013; Lazaridou et al., 2014). The setting later extends to a localization task, referred to as zero-shot object detection (ZSD) (Bansal et al., 2018; Zhu et al., 2019, 2020; Rahman et al., 2020). More recently, open-vocabulary object detection (OVD) (Zareian et al., 2021; Gu et al., 2022; Du et al., 2022b; Minderer et al., 2022) combines ZSD with weakly supervised object detection (WSD) to address the unrealistic constrain of traditional zero-shot settings. OVD assumes the availability of coarse-grained image-caption pairs, and attempts to generalize from limited fine-grained annotation of object categories to unseen ones. Nevertheless, this line of work positions words as object categories and isolates them from their linguistic context (*e.g.*, sentences). Our setup instead challenges models to perform language modeling in human-generated captions.

## 6 Conclusion and Future Work

The connection between language and their referents captures the grounded meaning of words, and an explicit treatment is key to empowering efficient open-world language learning abilities in humans and AI agents. This work introduces Grounded Open Vocabulary Acquisition (`GOVA`), a scalable formulation to examine grounding and fast mapping in open-world grounded language learning. We propose World-to-Words (`W2W`), a novel visually grounded language model to investigate a paradigm where the model initially acquires grounding ability during pre-training and subsequently applies this ability to quickly learn new words without explicit grounding supervision. Our empirical findings highlight the significance of visual grounding in neural word acquisition. Especially, we find that pre-trained `W2W` can serve as a foundation for fast mapping of novel grounded words via few-shot learning. We also conduct a comprehensive analysis to explore potential predictors influencing the performance of vision-language models, revealing both consistent and surprising behaviors with respect to human language learning patterns. These insights pave the way for future research in grounded language learning in the open world.

## Limitations

この研究では、言語が出来事、属性、マナー、精神状態などを根拠づけることができることを無視したオブジェクト中心の根拠付けに限定している。特にADV、NUM、VERB、PRONといった根拠となる単語の根拠となる意味は、バウンディングボックスだけでは十分に捉えることができない。今後は、その根拠となる意味の獲得を研究するために、より良いタスクの定式化を模索する必要がある。この線に沿った将来のエキサイティングな研究は、画像からビデオや環境との物理的相互作用に設定を拡張し、言語習得のために世界の豊かな時間的ダイナミクスを取り入れることである。さらに、我々は言語学習の社会的側面を無視し、子どもはコミュニケーションを通じて養育者から単語の参照元を推測する(Carpenter et al., 1998; Bloom, 2000)。また、今後の研究では、自然な対話から得られる根拠ある単語の獲得についても調査することができる。

## Ethics Statement

このプロジェクトは、人間による教科研究を通じて生み出された研究成果物を伴わない。W2Wは大きな可能性を持っているが、その倫理的・社会的影響を検証することは極めて重要である。計算モデルは事前に学習された言語モデルと広範なテキスト-画像データセットに依存しており、アルゴリズム内の公平性に問題をもたらす可能性のある隠れたバイアスを含む可能性がある。これらの意味を認識し、積極的に対処することで、将来的に言語学習エージェントとしてモデルが展開される場合、実務家の間での認識を高めることを目指す。

## Acknowledgments

この研究は、NSF IIS1949634, NSF SES-2128623 およびミシガン大学の Automotive Research Center (ARC) から一部支援を受けたものである。また、貴重なご意見をいただいた匿名査読者の方々に感謝いたします。

## References

アンカン・バンサル、カラン・シッカ、ガウラヴ・シャルマ、ラマ・チェルッパ、アジェイ・ディヴァカラン。2018. ゼロショット物体検出。欧州コンピュータビジョン会議(ECCV)議事録、ページ384-400にて。

ジョナタン・ビスク、アリ・ホルツマン、ジェシー・トーマスン、ジェイコブ・アンドレアス、ヨシュア・ベンジオ、ジョイス・チャイ、ミレッラ・ラパタ、アンジェリキ・ラザリドゥ、ジョナサン・メイ、アレクサンドル・ニスネビッチ、ニコラ・ピント、ジョセフ・トゥリアン。2020. 経験ベース言語。自然言語処理における経験的方法(EMNLP)に関する2020年会議の議事録、ページ8718-8735、オンライン。Association for Computational Linguistics.

ポール・ブルーム. 2000. 子どもは言葉の意味をどのように学ぶか. MITプレス.

スーザン・キャリー. 1978. 単語学習者としての子ども. 言語理論と心理的現実.

スーザン・キャリー、エルザ・バートレット。1978. 1つの新しい単語を獲得する。子どもの言語発達に関する論文と報告書, 15:17-29.

ニコラ・カリオン、フランシスコ・マサ、ガブリエル・シナヴェ、ニコラ・ウズニエル、アレクサンダー・キルヨルフ、セルゲイ・ザゴルイコ。2020. 変換器を用いたエンドツーエンドのオブジェクト検出。コンピュータビジョンに関する欧州会議、ページ213-229で。Springer.

Malinda Carpenter, Katherine Nagell, Michael Tomasello, George Butterworth, and Chris Moore. 1998. 生後9ヶ月から15ヶ月までの社会的認知、共同注意、コミュニケーション能力。子どもの発達の研究のための社会のモノグラフ、ページi-174。

サンチャゴ・カストロ、王羅雄、黄景秀、イアン・スチュワート、オアナ・イグナ、南劉、ジョナサン・ストロー、ラダ・ミハルチャ. 2022. ファイバー Fill-in-the-blanks as a challenge video understanding evaluation framework. 第60回計算言語学会年次大会予稿集(第1巻:ロングペーパー), ページ2925-2940.

Tyler A Chang and Benjamin K Bergen. 2022. ニューラル言語モデルにおける単語の獲得。計算言語学会論文誌, 10:1- 16.

陳陽、李林傑、于立、Ahmed El Kholy、Faisal Ahmed、Zhe Gan、Yu Cheng、劉景景景景。2020. Uniter: ユニバーサル画像テキスト表現学習。ECCVにて。

Grzegorz Chrupała, Ákos Kádár, and Afra Alishahi. 2015. 絵を通して言語を学ぶ。第53回計算言語学会年次大会および第7回自然言語処理国際合同会議(第2巻:短報)予稿集、ページ112-118、北京、中国。Association for Computational Linguistics(計算言語学会)。

Eve V Clark. 1995. 取得における語彙. 65. ケンブリッジ大学出版局.

マックス・コルティハート 1981. mrc psycholinguistic database. 実験心理学季刊誌 A, 33(4):497-505.

Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun P

## Limitations

In this work, we limit ourselves to object-centric grounding, which ignored that language can ground events, attributes, manners, mental states, etc. The grounded meaning of some groundable words, especially ADVs, NUMs, VERBs, and PRONs, cannot be fully captured by the bounding boxes alone. Future work should explore better task formulations to study the acquisition of their grounded meanings. An exciting future work along this line is to extend the setting from images to videos and physical interactions with the environment, and to incorporate the rich temporal dynamics of the world for language acquisition. In addition, we ignored the social aspects of language learning, where children infer the referents of words from their caregivers through communication (Carpenter et al., 1998; Bloom, 2000). Future work could also investigate grounded word acquisition from natural dialogue.

## Ethics Statement

This project does not involve any research artifacts generated through human subject studies. Despite the considerable promise of W2W, it is crucial to examine its ethical and societal implications. The computational model relies on pre-trained language models and extensive text-image datasets, which could contain hidden biases that may result in fairness problems within the algorithms. By recognizing and actively addressing these implications, we aim to increase awareness among practitioners if the model is deployed as a language-learning agent in the future.

## Acknowledgments

This work was supported in part by NSF IIS-1949634, NSF SES-2128623, and by the Automotive Research Center (ARC) at the University of Michigan. The authors would like to thank the anonymous reviewers for their valuable feedback.

## References

Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. 2018. Zero-shot object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 384–400.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.

Paul Bloom. 2000. *How children learn the meanings of words*. MIT press.

Susan Carey. 1978. The child as word learner. *Linguistic theory and psychological reality*.

Susan Carey and Elsa Bartlett. 1978. Acquiring a single new word. *Papers and Reports on Child Language Development*, 15:17–29.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer.

Malinda Carpenter, Katherine Nagell, Michael Tomasello, George Butterworth, and Chris Moore. 1998. Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the society for research in child development*, pages i–174.

Santiago Castro, Ruoyao Wang, Pingxuan Huang, Ian Stewart, Oana Ignat, Nan Liu, Jonathan Stroud, and Rada Mihalcea. 2022. Fiber: Fill-in-the-blanks as a challenging video understanding evaluation framework. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2925–2940.

Tyler A Chang and Benjamin K Bergen. 2022. Word acquisition in neural language models. *Transactions of the Association for Computational Linguistics*, 10:1–16.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *ECCV*.

Grzegorz Chrupała, Ákos Kádár, and Afra Alishahi. 2015. Learning language through pictures. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 112–118, Beijing, China. Association for Computational Linguistics.

Eve V Clark. 1995. *The lexicon in acquisition*. 65. Cambridge University Press.

Max Coltheart. 1981. The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4):497–505.

Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. 2022. Coarse-to-fine vision-language pre-training

with fusion in the backbone. *arXiv preprint arXiv:2206.07643*.

eng, et al. 2022. 視覚言語の事前学習 Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. 2022a. 視覚言語の事前学習済みモデルに関するサーベイ。人工知能に関する第31回国際合同会議(IJCAI-22)の議事録、ページ5436-5443。人工知能組織に関する国際合同会議。サーベイトラック。

Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. 2022b. 視覚言語モデルによるオープンボキャブラリー物体検出のプロンプトの学習。また、このような場合にも、「俯瞰的な視点」を持つことが重要である。

Mohamed Elhoseiny, Babak Saleh, and Ahmed Elgammal. 2013. 分類器を書く。純粋なテキスト記述を用いたゼロショット学習。IEEE International Conference on Computer Vision の議事録、ページ 2584-2591 に掲載されています。

Afsaneh Fazly, Afra Alishahi, and Suzanne Stevenson. 2010. を用いた、場所横断的な単語学習の確率的計算モデル。

Andrea Fromee, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. Devise:深い視覚意味埋め込みモデル. 神経情報処理システムの進歩, 26.

Lila R Gleitman and Barbara Landau. 1994. 語彙の獲得. mit Press.

ロベルタ・ミヒニック・ゴリンコフ、キャサリン・ハーシュ=パセック、ロワ・ブルーム、リンダ・B・スミス、アマンダ・L・ウッドワード、マベラ・アクタル、マイケル・トマセロ、ジョージ・ホリヒ。2000. ワード学習者になるために。語彙習得に関する議論。オックスフォード大学出版局。

arrivesマン、ジョシュア・テネンバウム、マイケル・ブラック。2007. ベイズ型フレームワークによるクロスシチュエーション・ワードラーニング. 神経情報処理システムの進歩, 20.

郭秀也・林恒義・郭偉哲・崔燕. 2022. 視覚と言語知識の蒸留によるオープンボキャブラリー物体検出. In International Conference on Learning Representations.

セルジオ・グダドラマ、エリック・ロドナー、ケイト・サエンコ、寧・チャン、ライアン・ファレル、ジェフ・ドナヒュー、トレバー・ダーレル。2014. 開口部オブジェクトの検索。In Robotics: science and systems, volume 2, page 6.

Agrim Gupta, Piotr Dollar, and Ross Girshick. 2019. LVIS: 大規模語彙インスタンス分割のためのデータセット. IEEE Conference on Computer Vision and Pattern Recognition の議事録にて。

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. 画像認識のための深層残差学習。コンピュータビジョンとパターン認識に関するIEEE会議の議事録、ページ770- 778にて。

Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. 2016. 自然言語による物体検索。IEEE conference on computer vision and pattern recognition の議事録、ページ 4555-4564 にて。

Xisen Jin, Junyi Du, Arka Sadhu, Ram Nevatia, and Xiang Ren. 2020. 構成フレーズの視覚的根拠に基づく継続的な学習。自然言語処理における経験的方法(EMNLP)に関する2020年会議の議事録、ページ2018-2029、オンライン。Association for Computational Linguistics(計算言語学会)。

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. 2021. エンドツーエンドのマルチモーダル理解のためのMdetr変調検出。IEEE/CVF International Conference on Computer Visionの議事録、ページ1780-1790に掲載されています。

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: 自然なシーンの写真におけるオブジェクトを参照する。自然言語処理における経験的方法(EMNLP)2014年会議の議事録、ページ787-798、ドーハ、カタール。Association for Computational Linguistics(計算言語学会)。

ロナルド・ケンカー、マーク・マキュル、アンジェリーナ・アビティーノ、タイラー・ヘイズ、クリストファー・カナン。2018. ニューラルネットワークにおける壊滅的な忘却の測定。人工知能に関するAAAI会議の議事録、第32巻にて。

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: 畳み込みや領域監視を用いない視覚言語変換器. 機械学習の国際会議, ページ 5583-5594 に収録。PMLR

エミール・クラマー、キース・ヴァン・デマー。2019. 参照表現の計算機による生成。更新された調査。

Harold W Kuhn. 1955. 割り当て問題に対するハンガリアン法. 海軍研究兵站季刊誌, 2(1-2):83-97.

Angeliki Lazaridou, Elia Bruni, and Marco Baroni. 2014. これは分布意味論と視覚世界の間のwampimuk? クロスモーダルマッピングなのか。第52回計算言語学会年次大会(第1巻:Long Papers)予稿集、ページ1403-1414。

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. 言語と視覚をマルチモーダルなスキップグラムモデルで結合する。Ang Li, Allan Jabri, Armand Joulin, and Laur

with fusion in the backbone. *arXiv preprint arXiv:2206.07643*.

Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. 2022a. A survey of vision-language pre-trained models. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5436–5443. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. 2022b. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14093.

Mohamed Elhoseiny, Babak Saleh, and Ahmed Elgammal. 2013. Write a classifier: Zero-shot learning using purely textual descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2584–2591.

Afsaneh Fazly, Afra Alishahi, and Suzanne Stevenson. 2010. A probabilistic computational model of cross-situational word learning. *Cognitive Science*, 34(6):1017–1063.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26.

Lila R Gleitman and Barbara Landau. 1994. *The acquisition of the lexicon*. mit Press.

Roberta Michnick Golinkoff, Kathryn Hirsh-Pasek, Lois Bloom, Linda B Smith, Amanda L Woodward, Nameera Akhtar, Michael Tomasello, and George Hollich. 2000. *Becoming a word learner: A debate on lexical acquisition*. Oxford University Press.

Noah Goodman, Joshua Tenenbaum, and Michael Black. 2007. A bayesian framework for cross-situational word-learning. *Advances in neural information processing systems*, 20.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. 2022. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*.

Sergio Guadarrama, Erik Rodner, Kate Saenko, Ning Zhang, Ryan Farrell, Jeff Donahue, and Trevor Darrell. 2014. Open-vocabulary object retrieval. In *Robotics: science and systems*, volume 2, page 6.

Agrim Gupta, Piotr Dollar, and Ross Girshick. 2019. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. 2016. Natural language object retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4555–4564.

Xisen Jin, Junyi Du, Arka Sadhu, Ram Nevatia, and Xiang Ren. 2020. Visually grounded continual learning of compositional phrases. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2018–2029, Online. Association for Computational Linguistics.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. 2021. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar. Association for Computational Linguistics.

Ronald Kemker, Marc McClure, Angelina Abitino, Tyler Hayes, and Christopher Kanan. 2018. Measuring catastrophic forgetting in neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

Emiel Krahmer and Kees van Deemter. 2019. Computational generation of referring expressions: An updated survey.

Harold W Kuhn. 1955. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Angeliki Lazaridou, Elia Bruni, and Marco Baroni. 2014. Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1403–1414.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining language and vision with a multimodal skip-gram model. In *Proceedings of the 2015 Conference of the North American Chapter of*

*the Association for Computational Linguistics: Human Language Technologies*, pages 153–163, Denver, Colorado. Association for Computational Linguistics.

ens Van Der Maatenの北米支部の2015年大会の議事録にて。2017. ウェブデータから視覚的n-gramを学習する。IEEE International Conference on Computer Visionの議事録、ページ4183-4192にて。

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. ビジュアルベルト。視覚と言語のためのシンプルでパフォーマントなベースライン. arXiv preprint arXiv:1908.03557.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. 2022年。接地言語-画像事前学習。IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10965- 10975 にて。

Changsong Liu, Lanbo She, Rui Fang, and Joyce Y. Chai. 2014. 協調的談話に基づく効率的な参照根拠のための確率的ラベリング。計算言語学会(Association for Computational Linguistics).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. ロベルタ:ロバスト最適化されたバート事前学習アプローチ.arXiv事前プリントarXiv:1907.11692.

呂玉江、朱Wanrong、王星、ミゲル・エクスタイン、王ウィリアム・ヤン。2022. Imaginationaugmented natural language understanding. 2022年、計算言語学会の北米支部大会の議事録。Human Language Technologies, pages 4392-4402, Seattle, United States. Association for Computational Linguistics(計算言語学会)。

Jacek Mandziuk and Lokendra Shastri. 1999. インクリメンタルクラス学習-長寿命でスケーラブルな学習へのアプローチ。IJCNN'99にて。ニューラルネットワークに関する国際合同会議. Proceedings (Cat. No. 99CH36339), volume 2, pages 1319-1324. IEEE.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. 2019. ニューロシンボリックなコンセプト学習器。自然監督からシーン、単語、文章を解釈する。学習表現に関する国際会議(ICLR).

Jiayuan Mao, Freda H. Shi, Jiajun Wu, Roger P. Levy, and Joshua B. Tenenbaum. 2021. 文法に基づく接地辞書学習。神経情報処理システムの進歩にて。

マオ・ジュワ、ファンタナ・ファンターン、アレキサンダー・トシェフ、オアナ・カンブル、アラン・L・ユイル、ケビン・マーフィー。2016. 曖昧さのない物体記述の生成と理解. IEEE conference on computer vision and pattern recognition, pages 11-20.

スティーブン・メリティ、カイミン・シオン、ジェームズ・ブラッドベリー、リチャード・ソッハー. 2017. ポインター・センチネル混合モデル。In International Conference on Learning Representations.

Antje S Meyer, Astrid M Sleiderink, and Willem JM Levelt. 1998. 物体を見たり名付けたりする。名詞句生成時の眼球運動。Cognition, 66(2):B25-B33.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al.2022年。ビジョン変換器を用いた単純なオープンボキャブラリーオブジェクト検出。欧州コンピュータビジョン会議にて。

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. LAMBADAデータセット。広範な談話コンテキストを必要とする単語予測。第54回計算言語学会年次大会(第1巻:ロングペーパー)予稿集、ページ1525-1534、ベルリン、ドイツ。Association for Computational Linguistics(計算言語学会)。

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. 知識ベースとしての言語モデル?自然言語処理における経験的方法に関する2019年会議と第9回自然言語処理に関する国際合同会議(EMNLP-IJCNLP)の議事録、ページ2463-2473、香港、中国。Association for Computational Linguistics(計算言語学会)。

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30kエンティティ。よりリッチな画像-文モデルのための領域-フレーズ対応の収集。IEEE国際コンピュータビジョン会議論文集, ページ2641-2649.

サラ・プラット、マーク・ヤツカール、ルカ・ワイス、アリ・ファルハディ、アニルダ・ケンバヴィ。2020. グラウンドド・コンテクスト・リコグニション。欧州コンピュータビジョン会議、ページ314-332にて。Springer.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many languages. 第58回計算言語学会年次大会予稿集: システムデモ、ページ101-108、オンライン。Association for Computational Linguistics(計算言語学会)。

*the Association for Computational Linguistics: Human Language Technologies*, pages 153–163, Denver, Colorado. Association for Computational Linguistics.

Ang Li, Allan Jabri, Armand Joulin, and Laurens Van Der Maaten. 2017. Learning visual n-grams from web data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4183–4192.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975.

Changsong Liu, Lanbo She, Rui Fang, and Joyce Y. Chai. 2014. Probabilistic labeling for efficient referential grounding based on collaborative discourse. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 13–18, Baltimore, Maryland. Association for Computational Linguistics.

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

Yujie Lu, Wanrong Zhu, Xin Wang, Miguel Eckstein, and William Yang Wang. 2022. Imagination-augmented natural language understanding. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4392–4402, Seattle, United States. Association for Computational Linguistics.

Jacek Mandziuk and Lokendra Shastri. 1999. Incremental class learning-an approach to longlife and scalable learning. In *IJCNN'99. International Joint Conference on Neural Networks. Proceedings (Cat. No. 99CH36339)*, volume 2, pages 1319–1324. IEEE.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. 2019. The neuro-symbolic concept learner: Interpreting scenes, words, sentences from natural supervision. *International Conference on Learning Representations (ICLR)*.

Jiayuan Mao, Freda H. Shi, Jiajun Wu, Roger P. Levy, and Joshua B. Tenenbaum. 2021. Grammar-based grounded lexicon learning. In *Advances in Neural Information Processing Systems*.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *International Conference on Learning Representations*.

Antje S Meyer, Astrid M Sleiderink, and Willem JM Levelt. 1998. Viewing and naming objects: Eye movements during noun phrase production. *Cognition*, 66(2):B25–B33.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. 2022. Simple open-vocabulary object detection with vision transformers. In *European Conference on Computer Vision*.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. 2020. Grounded situation recognition. In *European Conference on Computer Vision*, pages 314–332. Springer.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Shaolin Qu and Joyce Yue Chai. 2010. 仮想世界における状況対話のための文脈に基づく単語獲得。人工知能研究』37:247-277.

Shafin Rahman, Salman Khan, and Nick Barnes. 2020. ゼロショット物体検出のための視覚-意味アライメントの改善。人工知能に関するAAAI会議の議事録、第34巻、ページ 11932-11939。

テリー・レジェ. 2005. 単語の出現。形式と意味における注意の学習。認知科学, 29(6):819-865.

Hamid Rezatofighi, Nathan Tsoi, Jun Young Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 2019. ユニオン上の一般化された交差。バウンディングボックス回帰のためのメトリックと損失。IEEE/CVF conference on computer vision and pattern recognition, pages 658-666 にて。

ドブ・K・ロイ、アレックス・P・ペントランド。2002. 光景と音から単語を学習する。計算モデル. 認知科学, 26(1):113-146.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. マスク言語モデルのスコアリング. 計算言語学会の第58回年次大会の議事録、ページ2699-2712、オンライン。Association for Computational Linguistics(計算言語学会)。

Bharat Singh, Hengduo Li, Abhishek Sharma, and Larry S Davis. 2018. R-fcn-3000 at 30fps: デカップリング検出と分類. IEEE conference on computer vision and pattern recognition, pages 1081-1090 にて。

ジェフリー・マーク・シスキンド. 1996. 単語-意味マッピングを学習するための交差状況技法の計算論的研究。Cognition, 61(1-2):39-91.

リンダ・スミス、チェン・ユウ。2008. 幼児は、場所横断的な統計によって、単語参照マッピングを迅速に学習する。Cognition, 106(3):1558-1568.

リンダ・B・スミス、チェン・ユイ、アルフレド・ペレイラ。2007. From the outside-in: 幼児における身体化された注意。ヨーロッパ人工生命会議, ページ 445- 454 にて。Springer.

リチャード・ソッチャー、ミリンド・ガンジョ、クリストファー・D・マニング、アンドリュー・Ng. 2013. クロスモーダル転送によるゼロショット学習。神経情報処理システムの進歩, 26.

蘇偉江、朱秀州、曹裕恵、李ビン、呂レ偉、ウェイ古、戴知峰. 2020. Vl-bert: 汎用的な視覚言語表現の事前学習. In International Conference on Learning Representations.

Dıdac Surıs, Dave Epstein, Heng Ji, Shih-Fu Chang, and Carl Vondrick. 2020. 視覚シーンから単語を学習する。欧州コンピュータビジョン会議(ECCV).

Hao Tan and Mohit Bansal. 2020. Vokenization: 文脈に応じた視覚に基づく監視による言語理解の改善. 自然言語処理における経験的方法(EMNLP)に関する2020年会議論文集, ページ2066-2080, オンライン. Association for Computational Linguistics(計算言語学会)。

を、”enjoy ”と名付けた。1995. 音声言語理解における視覚情報と言語情報の統合。Science, 268(5217):1632- 1634.

王偉志、李東、鄭浩、宋浩宇、劉暁東、燕暁峰、高建峰、2002年。2022. 視覚的に拡張された言語モデリング.arXiv preprint arXiv:2205.10178.

Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. 2020. Phrasecut: における言語ベースの画像セグメンテーション. IEEE/CVF Conference on Computer Vision and Pattern Recognition の議事録、ページ 10216-10225.

徐飛、ジョシュア・B・テネンバウム。2007. ベイズ推論としての単語学習。心理学評論, 114(2):245.

Peter Young, Alice Lai, Mhakeh Hodosh, and Julia Hockenmaier. 2014. 画像記述から視覚的表記へ。イベント記述に対する意味推論のための新しい類似性メトリックス。計算言語学会論文誌, 2:67-78.

Chen Yu. 2005. 語彙獲得と物体分類の間のリンクの出現。計算機による研究。コネクションサイエンス, 17(3-4):381-397.

チェン・ユー、ダナ・H・バラード。2007. A unified model of early word learning: 統計的手がかりと社会的手がかりの統合. ニューロコンピューティング, 70(13-15):2149-2165.

Haonan Yu and Jeffrey Mark Siskind. 2013. 文で記述されたビデオからの接地言語学習。計算言語学会第51回年次大会(第1巻:Long Papers)予稿集, ページ53-63.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. 参照表現における文脈のモデル化。コンピュータビジョンに関する欧州会議、ページ69-85で。Springer.

Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. 2021. キャプションを用いたオープンボキャブラリーオブジェクト検出. IEEE/CVF Conference on Computer Vision and Pattern Recognition の議事録、ページ 14393-14402 に掲載されています。

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. 2022. リージョンクリップ。リージョンベースの言語-画像事前学習。IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16793-16803 にて。

Shaolin Qu and Joyce Yue Chai. 2010. Context-based word acquisition for situated dialogue in a virtual world. *Journal of Artificial Intelligence Research*, 37:247–277.

Shafin Rahman, Salman Khan, and Nick Barnes. 2020. Improved visual-semantic alignment for zero-shot object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11932–11939.

Terry Regier. 2005. The emergence of words: Attentional learning in form and meaning. *Cognitive science*, 29(6):819–865.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666.

Deb K Roy and Alex P Pentland. 2002. Learning words from sights and sounds: A computational model. *Cognitive science*, 26(1):113–146.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Bharat Singh, Hengduo Li, Abhishek Sharma, and Larry S Davis. 2018. R-fcn-3000 at 30fps: Decoupling detection and classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1081–1090.

Jeffrey Mark Siskind. 1996. A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61(1-2):39–91.

Linda Smith and Chen Yu. 2008. Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3):1558–1568.

Linda B Smith, Chen Yu, and Alfredo Pereira. 2007. From the outside-in: Embodied attention in toddlers. In *European Conference on Artificial Life*, pages 445–454. Springer.

Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. 2013. Zero-shot learning through cross-modal transfer. *Advances in neural information processing systems*, 26.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

Dıdac Surıs, Dave Epstein, Heng Ji, Shih-Fu Chang, and Carl Vondrick. 2020. Learning to learn words from visual scenes. *European Conference on Computer Vision (ECCV)*.

Hao Tan and Mohit Bansal. 2020. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2066–2080, Online. Association for Computational Linguistics.

Michael K Tanenhaus, Michael J Spivey-Knowlton, Kathleen M Eberhard, and Julie C Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632–1634.

Weizhi Wang, Li Dong, Hao Cheng, Haoyu Song, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. 2022. Visually-augmented language modeling. *arXiv preprint arXiv:2205.10178*.

Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. 2020. Phrasecut: Language-based image segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10216–10225.

Fei Xu and Joshua B Tenenbaum. 2007. Word learning as bayesian inference. *Psychological review*, 114(2):245.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Chen Yu. 2005. The emergence of links between lexical acquisition and object categorization: A computational study. *Connection science*, 17(3-4):381–397.

Chen Yu and Dana H Ballard. 2007. A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, 70(13-15):2149–2165.

Haonan Yu and Jeffrey Mark Siskind. 2013. Grounded language learning from video described with sentences. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 53–63.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer.

Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. 2021. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. 2022. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803.

朱鵬海、王漢暁、Saligrama Venkatesh. 2019. ゼロショット検出. IEEE Transactions on Circuits and Systems for Video Technology, 30(4):998-1010.

Pengkai Zhu, Hanxiao Wang, and Venkatesh Saligrama. 2020. Don't even onely look: ゼロショット検出のための特徴量の合成. IEEE/CVF Conference on Computer Vision and Pattern Recognition の議事録、ページ 11693-11702 に掲載されています。

Pengkai Zhu, Hanxiao Wang, and Venkatesh Saligrama. 2019. Zero shot detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(4):998–1010.

Pengkai Zhu, Hanxiao Wang, and Venkatesh Saligrama. 2020. Don't even look once: Synthesizing features for zero-shot detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11693–11702.

# A GOVA Dataset Details

## A.1 設定のイメージ図による比較

図8は、言語接地と接地言語学習に関連するタスクの定式化の比較を示したものである。これらのタスクの定式化のうち、我々のGVA(Grounded Open Vocabulary Acquisition)タスクは、視覚言語システムが視覚的に接地し、オブジェクト中心の言語モデリングを行うことに挑戦する唯一のタスクである。この定式化は自然でシンプルであり、マスク言語モデリングとオブジェクトローカライゼーションを行うための計算モデルに対する基本的な要求があるため、特にゼロショット解析に適している。

## A.2 Evaluation Protocols Explained

本論文では、グラウンデッドワード取得のための適切な評価プロトコルを提示する。このセクションでは、再現性のために、メトリクスと実装の詳細について、より詳細な説明を行う。

パープレキシティメトリックの詳細 我々は、単語wのパープレキシティを評価するために、クローステスト(Salazarら、2020;Jinら、2020)の先行慣行に従います。

$$\log \mathrm{PPL}(w) = -\log P(w|x_{\mathrm{img}}, x_{\mathrm{cap}})$$

しかし、言語モデルの大部分はテキストをセグメント化し、エンコードするためにサブワードトークン化法を採用している。特に、1つの語彙を複数のトークンに分割することができ、異なるトークナイザーは同じ入力に対して異なるトークンを導くことができる。そこで、我々は、トークン化器に依存したパープレキシティの尺度を導入する。トークン化器Tについては、単語w のN個のトークンをT（w）と表し

$$\log \mathrm{PPL}(w) = -\frac{1}{N} \sum_{t \in T(w)} \log P(t|x_{\mathrm{img}}, x_{\mathrm{cap}})$$

IoU メトリックの詳細 我々はKamathら(2021)と同じ課題に直面し、マスクされた単語に対して複数の参照語が可能である。同様に、Any-ProtocolとAllProtocolを採用し、接地検出タスクを評価する。n個の基底真理バウンディングボックスB = b 1, b 2,＿＿, b n_{}} とm個の予測バウンディングボックスB = b, b,, b 交差オーバーを仮定する。union e (IoU)^{{} e under−− Any-Protocols f^{}.} は、各グランドトゥルースオブジェクトに対して最もマッチする予測バウンディングボックスの平均IoUとして定義されます。

$$\mathrm{IoU}_{\mathrm{any}} = \frac{1}{n} \sum_{i \in \{1,2,\cdots,n\}} \max_{j \in \{1,2,\cdots,m\}} \mathrm{IoU}(b_i, \widetilde{b_j})$$

AllProtocolsの下での交差結合(IoU)は、グランドトゥルースのジョイントバウンディングボックスと予測バウンディングボックスの間のIoUとして定義される。

$$\mathrm{IoU}_{\mathrm{all}} = \mathrm{IoU}(\cup B, \cup \widetilde{B})$$

## A.3 Word List

- 見たままのセットには60語が含まれ、それぞれ80%のテストケースが用意されています:赤ちゃん、ボール、ビーチ、ベンチ、バイク、黒、金髪、青、男の子、ブラウン、建物、車、子供、暗い、犬、ドレス、顔、女性、フィールド、床、食品、少女、グラス、草、グレー、緑、ギター、ガイ、髪の毛、手、帽子、頭、馬、ジャケット、ジーンズ、女性、大、小、長、男性、オレンジ、パンツ、人、選手、赤、シャツ、歩道、サイン、小、雪、通り、ストライプ、テーブル、トップ、壁、水、白、女性、黄、若い。

- 31個の単語が未見集合に含まれ、それぞれ50個のテストケース[2] : 加齢、竹、裸足、ブラシ、ボタン、カフェ、チーズ、円形、教室、横断歩道、多様、医師、ロバ、象、ふわふわ、異物、体育館、心臓、新生、パン、ピザ、製品、セキュリティ、シンク、スター、ストーブ、生徒、教師、電話、暖かさ。

# B Computational Model Details

## B.1 Pre-training Objectives

マスク言語モデリング(MLM)。MLMの頭部は複数の場所に配置することができ、我々の設計は小規模な学習に関する予備実験の後の探求である。RoBERTaの設定に厳密に従い、huggingface[3]の実装に基づき、2層MLPでMLMヘッドを実装する。 接地可能なフレーズの単語は0.4の確率で、接地不可能な領域の単語は0.1の確率でマスクされる。マスクするトークンは、RoBERTaに従って、80%の確率でMASKに、10%の確率でランダムトークンに、10%の確率で何もしないように割り当てる。

[2] a few words (product, steep, telephone) has one less test case due to the availability of Flickr30K Entities Dataset.

[3] https://huggingface.co/docs/transformers/model_doc/roberta

# A GOVA Dataset Details

## A.1 Illustrated Comparison of Setting

We present an illustrated comparison of task formulations related to language grounding and grounded language learning in Figure 8. Among these task formulations, our Grounded Open Vocabulary Acquisition (GOVA) task is the only one that challenges vision-language systems to perform visually grounded and object-centric language modeling. The formulation is natural and simple, with fundamental requirements on computational models to perform masked language modeling and object localization, and thus is particularly good for zero-shot analysis.

## A.2 Evaluation Protocols Explained

We present an adequate evaluation protocol for *grounded* word acquisition in the main paper. This section provides more in-depth explanation for the metrics and implementation details for reproducibility purposes.

**Perplexity Metric Details** We follow prior practice in cloze tests (Salazar et al., 2020; Jin et al., 2020) to evaluate the perplexity of a word $w$. We use log pseudo-perplexity in masked language modeling, defined as

$$\log \text{PPL}(w) = -\log P(w|x_{\text{img}}, x_{\text{cap}})$$

However, the majority of the language models employ sub-word tokenization methods to segment and encode text. In particular, one lexical word can be segmented into several tokens, and different tokenizers can lead to different tokens for the same input. We thus introduce a tokenizer-dependent measure for perplexity. For tokenizer $T$, we represent the $N$ tokens of word $w$ as $T(w)$ and

$$\log \text{PPL}(w) = -\frac{1}{N} \sum_{t \in T(w)} \log P(t|x_{\text{img}}, x_{\text{cap}})$$

**IoU Metric Details** we face the same challenge as Kamath et al. (2021) where multiple referents are possible for a masked word. In a similar manner, we adopt the Any-Protocol and All-Protocol to evaluate the grounded detection task. Assuming $n$ ground truth bounding boxes $B = \{b_1, b_2, \cdots, b_n\}$ and $m$ predicted bounding boxes $\widetilde{B} = \{\widetilde{b_1}, \widetilde{b_2}, \cdots, \widetilde{b_m}\}$. The intersection-over-union (IoU) under Any-Protocols is defined as the average IoU of the best matching predicted bounding box for each ground truth object:

$$\text{IoU}_{\text{any}} = \frac{1}{n} \sum_{i \in \{1,2,\cdots,n\}} \max_{j \in \{1,2,\cdots,m\}} \text{IoU}(b_i, \widetilde{b_j})$$

The intersection-over-union (IoU) under All-Protocols is defined as the IoU between the joint bounding box of ground truth and predicted bounding boxes:

$$\text{IoU}_{\text{all}} = \text{IoU}(\cup B, \cup \widetilde{B})$$

## A.3 Word List

- 60 words are in the seen-set, each with 80 test cases: baby, ball, beach, bench, bike, black, blond, blue, boy, brown, building, car, child, dark, dog, dress, face, female, field, floor, food, girl, glasses, grass, gray, green, guitar, guy, hair, hand, hat, head, horse, jacket, jeans, lady, large, little, long, man, orange, pants, person, player, red, shirt, sidewalk, sign, small, snow, street, striped, table, top, wall, water, white, woman, yellow, young.

- 31 words are in the unseen-set, each with 50 test cases[2]: aged, bamboo, barefoot, brush, button, cafe, cheese, circular, classroom, crosswalk, diverse, doctor, donkey, elephant, fluffy, foreign, gym, heart, newborn, pan, pizza, product, security, sink, star, steep, stove, student, teacher, telephone, warm.

# B Computational Model Details

## B.1 Pre-training Objectives

**Masked Language Modeling (MLM).** The MLM head can be placed at multiple possible places, and our design is an exploration after preliminary experiments on smaller-scale training. We strictly follow the setup of RoBERTa to implement the MLM head with a two-layer MLP, based on the implementation of huggingface[3]. Words in groundable phrases are masked with a probability of 0.4 and those in non-groundable regions are masked with a lower probability of 0.1. For a token selected to mask, we follow RoBERTa to assign a probability of 80% to replace with MASK, 10% with a random token, and 10% to do nothing.

[2] a few words (product, steep, telephone) has one less test case due to the availability of Flickr30K Entities Dataset.

[3] https://huggingface.co/docs/transformers/model_doc/roberta


図8:接地言語学習に関連するタスクの定式化の比較を示す図。

| Tasks (Inference Time) | Language Input | Visual Input | Language Output | Vision Output | Example Dataset(s) |
|---|---|---|---|---|---|
| マスク言語モデリング 知識プローピング 読解力 | クローンテスト テストコンテキスト 、クローンテスト | -<br>-<br>- | 欠落した単語<br>欠落した単語<br>欠落した単語 | -<br>-<br>- | WikiText-103 (Merity et al., 2017)<br>LAMA (Petroni et al., 2019)<br>LAMBADA (Paperno et al., 2016) |
| 画像キャプション 塗りつぶし-塗りつぶし VQA 視覚的マスク言語モデリング | -<br>クローンテスト、(選択)クローンテスト | Image<br>画像/映像画像、バウンディングボックス | Caption<br>テキスト欠落(選択)<br>単語欠落 | -<br>-<br>- | Flickr30k (Young et al., 2014)<br>FIBER (Castro et al., 2022)<br>VisCOLL (Jin et al., 2020) |
| 物体検索参照式理解法 | Referring Expression<br>Referring Expression | Image, Bounding Boxes<br>Image | -<br>- | Bounding Boxes<br>Bounding Boxes | ReferIt (Kazemzadeh et al., 2014),<br>RefCOCO* (Yu et al., 2016; Mao et al., 2016) |
| Phrase Grounding | Caption, Referring Expressions | Image | - | Bounding Boxes | Flickr30K Entities (Plummer et al., 2015) |
| 物体検出 ゼロショット物体検出<br>オープンボキャブラリ物体検出 | で見たクラス 未見のクラス 事前学習用語彙 | Image<br>Image<br>Image | Classes<br>Classes<br>Words | バウンディングボックス バウンディングボックス バウンディングボックス | LVIS (Gupta et al., 2019) |
| Grounded Open Vocabulary Acquisition | Cloze Test | Image | Missing Word | Bounding Boxes | **GOVA (Ours)** |

表 5: 接地言語学習に関連するタスクの定式化の比較。

オブジェクトローカライゼーション(OL)。我々はMDETRに従って、3層MLPでオブジェクト埋め込みをデコードし、バウンディングボックスを生成する。ほとんどの先行研究と同様に、我々は信頼度が0.7以下のボックスに対してフィルタを適用する。我々のフレームワークでは、これはオブジェクトがオブジェクトなしラベル*(図4)に対応し、0.3を超える確率を持つことを意味する。我々はDETRに厳密に従い、提案されたボックスとハンガリー語の損失を持つグランドトゥルースボックスとの間の二分割マッチングを実行する。予測されたボックスは、一般化交差結合(GIoU)損失とL1損失によって、グランドトゥルースに向かって最適化される。

接地 位置合わせでは、モデルは各オブジェクト表現を文中のトークンに257の固定長で対応付けることを学習するが、これはおそらくMASKまたは追加のオブジェクトなしラベル*である可能性がある(図4)。オブジェクトとトークンは0.1以上のマッピング確率が与えられれば一致するとみなされる。我々は完全連結層を用いて、クロスエントロピー損失でトークン位置の分布を予測する。意味的アライメントでは、モデルは単語埋め込みを接地先のオブジェクト埋め込みに近づけ、関係のないペアを遠くに押し出すように学習する。この目的のために、全てのオブジェクトと接地可能なトークンについてMDETRで定義された対照的損失関数を厳密に踏襲する。

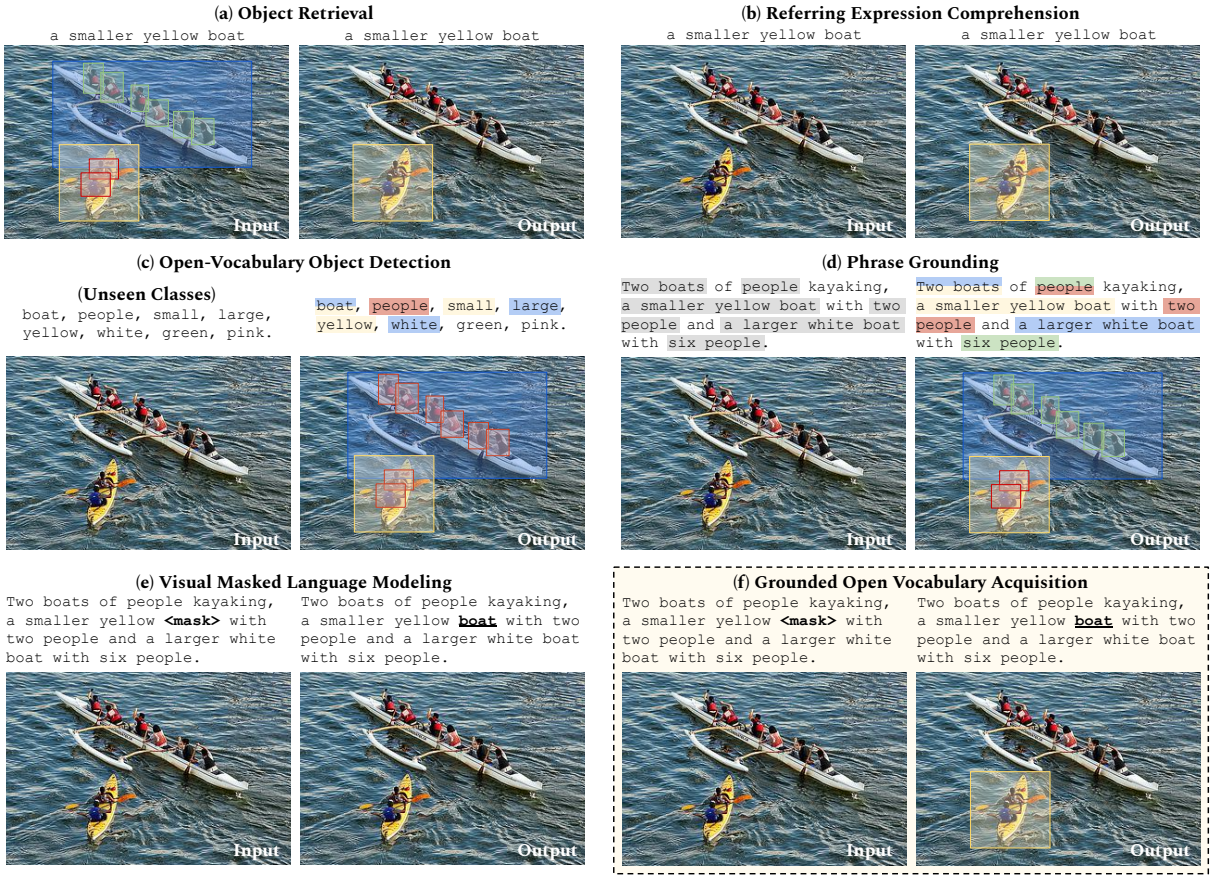



Figure 8: An illustrated comparison of task formulations related to grounded language learning.

| Tasks (Inference Time) | Language Input | Visual Input | Language Output | Vision Output | Example Dataset(s) |
|---|---|---|---|---|---|
| Masked Language Modeling | Cloze Test | - | Missing Word | - | WikiText-103 (Merity et al., 2017) |
| Knowledge Probing | Cloze Test | - | Missing Word | - | LAMA (Petroni et al., 2019) |
| Reading Comprehension | Context, Cloze Test | - | Missing Word | - | LAMBADA (Paperno et al., 2016) |
| Image Captioning | - | Image | Caption | - | Flickr30k (Young et al., 2014) |
| Fill-in-the-Blank VQA | Cloze Test, (Choices) | Image/Video | Missing Text (Choice) | - | FIBER (Castro et al., 2022) |
| Visual Masked Language Modeling | Cloze Test | Image, Bounding Boxes | Missing Word | - | VisCOLL (Jin et al., 2020) |
| Object Retrieval | Referring Expression | Image, Bounding Boxes | - | Bounding Boxes | ReferIt (Kazemzadeh et al., 2014), |
| Referring Expression Comprehension | Referring Expression | Image | - | Bounding Boxes | RefCOCO* (Yu et al., 2016; Mao et al., 2016) |
| Phrase Grounding | Caption, Referring Expressions | Image | - | Bounding Boxes | Flickr30K Entities (Plummer et al., 2015) |
| Object Detection | Seen Classes | Image | Classes | Bounding Boxes | |
| Zero-Shot Object Detection | Unseen Classes | Image | Classes | Bounding Boxes | LVIS (Gupta et al., 2019) |
| Open-Vocabulary Object Detection | Pre-training Vocabulary | Image | Words | Bounding Boxes | |
| Grounded Open Vocabulary Acquisition | Cloze Test | Image | Missing Word | Bounding Boxes | **GOVA (Ours)** |

Table 5: Comparison of task formulations related to grounded language learning.

**Object Localization (OL).** We follow MDETR to decode object embeddings with a three-layer MLP to produce bounding boxes. Similar to most prior work, we apply a filter over boxes with confidence below 0.7. In our framework, this means that the object corresponds to the no-object label $\varnothing$ (Figure 4) with a probability over 0.3. We strictly follow DETR to perform bipartite matching between proposed boxes and ground truth boxes with a Hungarian loss. The predicted boxes are optimized towards ground truth by the generalized intersection-over-union (GIoU) loss and the L1 loss.

**Grounding.** In positional alignment, the model learns to map each object representation to tokens in the sentence with a fixed length of 257, which could possibly be a MASK or an additional no-object label $\varnothing$ (Figure 4). The object and the token are considered a match given a mapping probability over 0.1. We use a fully-connected layer to predict the distribution over token positions with cross-entropy loss. In semantic alignment, the model learns to bring word embeddings closer to the object embeddings that they ground to, and push the unrelated pairs farther. We strictly follow the contrastive loss function defined in MDETR for every object and groundable token for this purpose.

### B.2 Few-shot Learning Details.

バウンディングボックスや単語-オブジェクトマッピングのアノテーションがないため、少数サンプルの新しい単語学習において、マスク言語モデリング(MLM)のみでW2Wを学習させる。少ないサンプル数を考慮し、バッチサイズを8に減らし、収束基準を固定数、すなわち50ステップに設定する。残りの実験設定は全て事前学習と同じである。

## C Experiment Reproducibility

### C.1 W2W Implementation Details

我々のW2Wモデルは主に、画像とテキスト領域からのユニモーダルエンコーダからの入力を持つ1つのクロスモーダル変換器から構成されている。特に、画像エンコーダとしてTIMM[4]からImageNetで事前学習したResNet-50(He et al., 2016)、テキストエンコーダとしてhuggingface[5]からのRoBERTa-base(Liu et al., 2019)を選択する。クロスモーダルエンコーダと2つのデコーダはそれぞれ、8つの注意ヘッド、入力と出力の次元512、および内層の次元2,048を持つ4つのトランスフォーマーブロックから構成されています。さらに、50の学習可能なオブジェクトクエリが含まれ、クロスモーダルデコーダに問い合わせ、バウンディングボックス提案を生成する。

### C.2 Hyper-parameter Decisions

再現性を高めるため、主要なハイパーパラメータチューニングの決定事項を記載しています。詳細は補足コードを参照してください。

- Learning Rate:
  - Image Encoder: frozen
  - Text Encoder: $1 \times 10^{-5}$
  - Multi-modal Transformer: $1 \times 10^{-4}$
- Batch Size: 128
- Pre-training Loss Coefficients:

  - MLM Loss: 32 - 位置合わせのためのクロスエントロピー。1 - 意味アライメントのための対照的損失。1 - L1 ローカライゼーションロス: 5 - GIoU ローカライゼーションロス: 2

- Few-shot Learning:

  - バッチサイズ: 8 - その他のハイパーパラメータ。事前学習と同じ

[4] https://github.com/rwightman/pytorch-image-models

[5] https://huggingface.co/docs/transformers/model_doc/roberta

### C.3 Computational Resources

W2Wモデルは8台のNVidia A40 GPUで事前学習しています。混合精度の事前学習とバッチサイズ128で、W2Wは15万ステップ、各ステップに約1.4秒を要する学習を行いました。

### C.4 Evaluation on GOVA

W2W 提案するW2Wモデルでは、GOVAテストが与えられ、それに対応する画像とテキストのクローンペアがモデルに渡される場合、バウンディングボックス予測は、クローン内の少なくとも一つのマスクされたトークンにマッピングされたバウンディングボックス提案のみを保持することによって生成され、マスクされたトークン予測結果は、その言語モデリングヘッドから直接デコードされる。

VisualBERT 「Detect-and-Recognize」ベースラインモデルVisualBERTでは、VisualBERTのフレーズ接地微調整版を用いて物体位置決めを行い、言語モデリングヘッドを持たないため、別のバニラ事前学習したVisualBERTを用いてマスクトークン予測を行う。具体的には、バウンディングボックスローカライゼーション部分については、標準的なフレーズ接地タスクとして扱い、(Li et al., 2019)に従って、最後のマスクトークンにおけるトップ1バウンディングボックス予測を出力として選択します。

ViLT+MDETR "Produce-and-Localize "ベースラインモデルViLT+MDETRでは、ステージ1で入力画像とテキストをViLTに送り、そのトップ1のクローントークン予測結果を収集する。次に、ステージ2において、入力画像とViLTで完成したテキストをMDETRに入力し、フレーズグラウンディングを行い、元のクローンに関連するオブジェクトをローカライズする。最後に、ViLTによるクローントークン予測結果を、MDETRによるバウンディングボックス提案とともに、GOVA評価に利用する。

## D Addendum to Results

### D.1 Ablation Study

W2Wモデルの有効性をピンポイントで確認するために、いくつかのW2Wモデルバリエーションでアブレーション研究を行った。その中には、言語エンコーダの初期化なし(w/o Init)、接地目的なし(w/o G)、物体中心表現なし(w/o O)、視覚入力なしのテキストのみの設定(w/o V)などが含まれる。一貫性を持たせるために、各バリエーションにおける変換層の数とパラメータの数を制御する。

### B.2 Few-shot Learning Details.

Since no bounding box or word-object mappings annotation is available, we train W2W with only masked language modeling (MLM) in few-sample new word learning. We reduce the batch size to 8 considering the fewer number of samples, and set the convergence criteria to a fixed number, *i.e.*, 50 steps. All the rest of the experimental settings remain the same as pre-training.

## C Experiment Reproducibility

### C.1 W2W Implementation Details

Our W2W model mainly consists of one cross-modal transformer with inputs from uni-modal encoders from image and text domain. Specially, we select the ResNet-50 (He et al., 2016) pre-trained on ImageNet from TIMM[4] as the image encoder, and RoBERTa-base (Liu et al., 2019) from huggingface[5] as the text encoder. The cross-modal encoder and two decoders each consists of 4 transformer blocks with 8 attention heads, an input and output dimensionality of 512, and an inner-layer dimensionality of 2,048. Besides, 50 learnable object queries are included to query the cross-modal decoder to generate bounding box proposals.

### C.2 Hyper-parameter Decisions

We include the major hyper-parameter tuning decisions for reproducibility purpose. For more details, please refer to the supplementary codes.

- Learning Rate:
  - Image Encoder: frozen
  - Text Encoder: $1 \times 10^{-5}$
  - Multi-modal Transformer: $1 \times 10^{-4}$
- Batch Size: 128
- Pre-training Loss Coefficients:
  - MLM Loss: 32
  - Cross Entropy for Positional Alignment: 1
  - Contrastive Loss for Semantic Alignment: 1
  - L1 Localization Loss: 5
  - GIoU Localization Loss: 2
- Few-shot Learning:
  - Batch size: 8
  - Other Hyper-parameters: Same as Pre-training

[4] https://github.com/rwightman/pytorch-image-models

[5] https://huggingface.co/docs/transformers/model_doc/roberta

### C.3 Computational Resources

Our W2W models is pre-trained on 8 NVidia A40 GPUs. With mixed-precision pre-training and a batch size of 128, W2W was trained for 150,000 steps where each step takes about 1.4 second.

### C.4 Evaluation on GOVA

**W2W** For our proposed W2W model, given a GOVA test, with its corresponding image and textual cloze pair passing into the model, the bounding box predictions are generated by keeping only the bounding box proposals that are mapped to at least one masked token within the cloze, while the masked token prediction results are directly decoded from its language modeling head.

**VisualBERT** For the "Detect-and-Recognize" baseline model VisualBERT, we use phrase-grounding fine-tuned version of VisualBERT to perform object localization, and, as it lacks the language modeling head, another vanilla pre-trained VisualBERT to perform mask token prediction. Specifically, for the bounding box localization part, we treat it as a standard phrase grounding task and follow (Li et al., 2019) to select the top-1 bounding box prediction in the last masked token as the output.

**ViLT+MDETR** For the "Produce-and-Localize" baseline model ViLT + MDETR, in stage one, we feed the input image and text into ViLT, collecting its top-1 cloze token prediction result. Then, at stage two, the input image and ViLT-completed text are fed into MDETR, performing phrase-grounding to localize the object associated with the original cloze. Finally, the cloze token prediction result from ViLT together with the bounding box proposals from MDETR are used for GOVA evaluation.

## D Addendum to Results

### D.1 Ablation Study

We performed an ablation study on several W2W model variants to pinpoint what makes our W2W model effective. These included models without language encoder initialization (w/o Init), without grounding objective (w/o G), without any object-centric representation (w/o O), and a text-only setup without any vision input (w/o V). For consistency, we control the number of transformer layers and the number of parameters for each variation. Despite tweaking various hyperparameters, no significant improvements were observed. As a result,

様々なハイパーパラメータを微調整したにもかかわらず、有意な改善は見られなかった。その結果、W2Wモデルと同じハイパーパラメータを保持した。

- w/o G: 3.2 節で既に述べたように、接地損失を伴わないモデルバリエーションを指す。

- w/o O: このモデルでは、マスク言語モデリング(MLM)目的のみを保持し、すべてのオブジェクト中心の表現を除外する。このモデルでは、オブジェクトデコーダ変換器は不要であるため、接地やローカライズは行わない。その代わりに、12個の変換器ブロックをすべてマルチモーダルエンコーダに統合し、MLMの目的語を直接付加する。

- w/o V: このテキストのみのモデルは、視覚の入力や監視なしに動作し、12個のトランスフォーマーブロックを追加したユニモーダル言語モデル(RoBERTa)に還元される。

Chang and Bergen (2022)のユニモーダル言語モデルにおける分析に続き、モデル予測値とユニグラム分布の間のKL-Divergenceを図9に示す。すぐに観察されるのは、すべての変種が事前学習の10 2ステップ程度で浅いunigram統計量に収束することである。これは、ユニモーダル言語モデルがより複雑な文脈表現を獲得する前にunigramに収束するというChang and Bergen (2022)の知見と一致する。我々は、MLMが唯一の事前学習目的であるテキストのみとW2W w/o Oの両方のケースにおいて、10 4ステップの学習でもモデルがunigram単語分布の周辺に留まる傾向があることに気付いた。しかし、オブジェクト中心の表現を持つ変種は、すぐにunigram分布から外れてしまう。一方、言語モデルの初期化を行ったモデルは、一字形分布から急速に離れ、目的を持ったモデルは、わずかに速い偏差を持つ。これらの結果は、視覚言語モデルが大規模なコーパスでユニモーダルな事前学習から利益を得ることができ、オブジェクト表現に対して言語モデリングを行うことが重要であることを確認するものである。我々は、モデルの動作を理解するために、単グラムからのKL-Divergenceを比較し、メトリック自体は、接地されたオープンボキャブラリーの取得におけるシステムの性能の評価として機能しないことに注意する。

### D.2 Addendum to Results in Multi-Class Incremental Learning

表 6 に追加結果を示す。

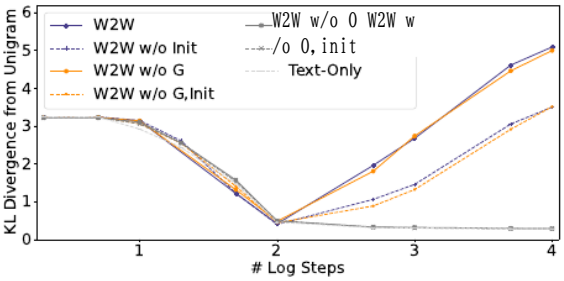



図9:モデルのトークン予測と学習コーパスの1gram分布の間のKL-divergence。

| # Samples | Seen log G-PPL$_{all}$ (↓) | | Unseen log G-PPL$_{all}$ (↓) | |
|---|---|---|---|---|
| | W2W | W2W$_{w/o\ G}$ | W2W | W2W$_{w/o\ G}$ |
| 0 | **1.79** | 2.33 | **11.58** | 11.89 |
| 8 | **3.15** | 3.63 | **3.09** | 3.32 |
| 16 | **3.36** | 3.76 | **2.64** | 2.85 |
| 24 | **3.05** | 3.46 | **2.07** | 2.67 |
| 32 | **3.07** | 3.62 | **2.01** | 2.54 |

表6:多クラス漸増学習における、サンプルサイズ8から32の範囲で各未見単語の対数G-PPL(All-Protocol)。

### D.3 Learning New Words through One-Class Incremental Learning.

さらに、単語に特化した1クラス漸増学習の設定で、より対照的な研究を行う。事前学習されたモデルは、$_{\triangle}V$ unseen$_{\triangle}$ = 1の数発の学習セッションから1つの未見単語を獲得するタスクを課される。本節の結果は、新しいセッションの直後のテストから得られたものである。テスト結果を表7に示す。ここでも、わずか8個のサンプルで、W2Wは満足に低い接地率を達成できることが観察される。ほとんどの場合、W2Wはグランドレスベースラインよりも未見語を獲得する能力が高いことがわかる。

we retained the same hyperparameters as in the W2W model.

- w/o G: This refers to the model variant without grounding loss, as has already been described in Section 3.2;
- w/o O: This variant excludes all object-centric representations, retaining only the masked language modeling (MLM) objective. With this model, the object decoder transformer is unnecessary, thus no grounding nor localization is performed. Instead, we consolidate all 12 transformer blocks into the multi-modal encoder and directly attach the MLM objective to it.
- w/o V: This text-only model operates without any vision input or supervision, reducing it to a unimodal language model (RoBERTa) with 12 additional transformer blocks.

Following the analysis of Chang and Bergen (2022) in unimodal language models, we present the KL-Divergence between the model predictions and the unigram distribution in Figure 9. An immediate observation is that all variants converge to the shallow unigram statistics at around $10^2$ steps of pre-training. This aligns with the findings of Chang and Bergen (2022) that unimodal language models would converge to unigram before acquiring more complicated contextual representations. We noticed that in both text-only and $\texttt{W2W}_{\text{w/o O}}$ cases where MLM is the only pre-training objective, the models tend to stay around the unigram word distribution even with $10^4$ steps of training. However, variants with an object-centric representation quickly departed from the unigram distribution. Comparatively, models with language model initialization moves quickly away from the unigram distribution, and models with a grounded objective have a marginally faster deviation. These results confirm that vision-language models can benefit from unimodal pre-training on a large corpus, and that performing language modeling upon object representations is crucial. We note that we compare the KL-Divergence from unigram only to understand the models' behaviors, and the metric itself does not serve as an evaluation of a system's performance in grounded open vocabulary acquisition.

### D.2 Addendum to Results in Multi-Class Incremental Learning

We present additional results in Table 6.

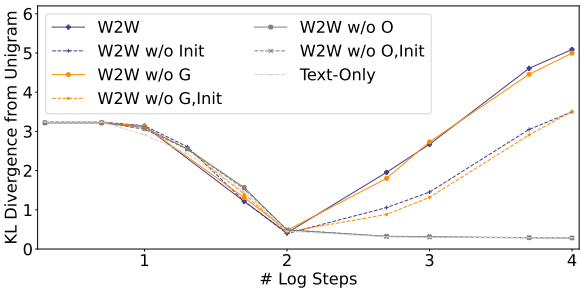



Figure 9: KL-divergence between model's token prediction and the unigram distribution of the training corpus.

| # Samples | Seen log G-PPL$_{\text{all}}$ (↓) | | Unseen log G-PPL$_{\text{all}}$ (↓) | |
|---|---|---|---|---|
| | W2W | $\text{W2W}_{\text{w/o G}}$ | W2W | $\text{W2W}_{\text{w/o G}}$ |
| 0 | **1.79** | 2.33 | **11.58** | 11.89 |
| 8 | **3.15** | 3.63 | **3.09** | 3.32 |
| 16 | **3.36** | 3.76 | **2.64** | 2.85 |
| 24 | **3.05** | 3.46 | **2.07** | 2.67 |
| 32 | **3.07** | 3.62 | **2.01** | 2.54 |

Table 6: The log G-PPL (All-Protocol) of seen and unseen words in multi-class incremental learning, each unseen word with a sample size ranging from 8 to 32.

### D.3 Learning New Words through One-Class Incremental Learning.

We further perform a more controlled study with a word-specific one-class incremental learning setting. The pre-trained model is tasked to acquire one single unseen word from a few-shot learning session with $|\mathcal{V}_{\text{unseen}}| = 1$. The results of this section are obtained from the test immediately following the new session. We present the test result in Table 7. Again, we observe that with as few as 8 samples, W2W can achieve a satisfyingly low grounded perplexity. In the majority of the cases, W2W demonstrates the better ability to acquire unseen words over the groundless baseline.

| # Samples | | 0 | 8 | 16 | 24 | 32 | # Samples | | 0 | 8 | 16 | 24 | 32 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| crosswalk | W2W | 10.82 | 8.48 | 7.43 | 7.70 | 5.95 | donkey | W2W | 8.70 | 0.84 | 0.81 | 0.67 | 0.79 |
| | $W2W_{w/o\ G}$ | 10.91 | 10.88 | 7.53 | 7.15 | 7.5 | | $W2W_{w/o\ G}$ | 9.69 | 1.97 | 1.99 | 2.35 | 2.01 |
| cheese | W2W | 12.16 | 2.62 | 3.00 | 1.27 | 1.04 | barefoot | W2W | 9.71 | 6.93 | 4.58 | 5.55 | 6.27 |
| | $W2W_{w/o\ G}$ | 13.07 | 2.81 | 3.13 | 2.56 | 1.49 | | $W2W_{w/o\ G}$ | 9.95 | 6.52 | 4.67 | 5.74 | 5.88 |
| star | W2W | 8.70 | 1.49 | 1.47 | 1.09 | 1.18 | elephant | W2W | 15.24 | 1.44 | 1.65 | 1.81 | 1.44 |
| | $W2W_{w/o\ G}$ | 10.59 | 2.93 | 2.10 | 1.99 | 1.39 | | $W2W_{w/o\ G}$ | 14.75 | 2.17 | 1.98 | 1.73 | 1.61 |
| classroom | W2W | 3.96 | 0.47 | 0.36 | 0.43 | 0.32 | heart | W2W | 9.34 | 2.97 | 1.90 | 1.76 | 1.76 |
| | $W2W_{w/o\ G}$ | 5.10 | 0.95 | 0.88 | 1.05 | 0.95 | | $W2W_{w/o\ G}$ | 9.31 | 2.99 | 2.50 | 2.65 | 2.96 |
| fluffy | W2W | 16.44 | 1.88 | 1.78 | 0.82 | 1.36 | gym | W2W | 5.13 | 2.14 | 0.44 | 0.74 | 0.69 |
| | $W2W_{w/o\ G}$ | 15.61 | 1.83 | 1.71 | 1.37 | 1.47 | | $W2W_{w/o\ G}$ | 4.88 | 3.73 | 1.30 | 1.08 | 1.45 |
| circular | W2W | 15.21 | 1.59 | 1.07 | 1.55 | 1.23 | security | W2W | 15.08 | 1.07 | 0.81 | 1.28 | 0.71 |
| | $W2W_{w/o\ G}$ | 15.12 | 2.25 | 2.25 | 1.81 | 1.61 | | $W2W_{w/o\ G}$ | 14.75 | 1.50 | 1.22 | 1.53 | 1.17 |
| sink | W2W | 14.23 | 1.17 | 0.92 | 1.11 | 1.38 | cafe | W2W | 6.28 | 1.90 | 1.38 | 1.98 | 1.39 |
| | $W2W_{w/o\ G}$ | 15.49 | 1.84 | 1.65 | 1.60 | 1.84 | | $W2W_{w/o\ G}$ | 7.03 | 2.17 | 1.92 | 2.08 | 1.72 |
| doctor | W2W | 13.03 | 1.17 | 1.05 | 1.38 | 1.18 | teacher | W2W | 16.68 | 1.95 | 2.15 | 1.52 | 1.48 |
| | $W2W_{w/o\ G}$ | 12.44 | 1.17 | 1.23 | 1.39 | 1.58 | | $W2W_{w/o\ G}$ | 16.08 | 2.68 | 2.37 | 1.85 | 1.83 |
| foreign | W2W | 9.48 | 0.62 | 0.95 | 0.85 | 0.47 | student | W2W | 16.28 | 1.38 | 1.07 | 1.20 | 1.03 |
| | $W2W_{w/o\ G}$ | 10.01 | 1.03 | 0.88 | 1.18 | 0.95 | | $W2W_{w/o\ G}$ | 16.52 | 2.21 | 1.29 | 1.40 | 1.61 |
| diverse | W2W | 16.44 | 0.60 | 0.22 | 0.52 | 0.24 | newborn | W2W | 16.43 | 1.71 | 0.88 | 0.91 | 1.11 |
| | $W2W_{w/o\ G}$ | 16.05 | 0.81 | 0.65 | 0.97 | 0.65 | | $W2W_{w/o\ G}$ | 16.30 | 2.02 | 1.32 | 1.61 | 1.76 |
| product | W2W | 10.25 | 0.84 | 0.75 | 1.39 | 1.15 | pan | W2W | 12.04 | 1.70 | 2.12 | 1.87 | 2.02 |
| | $W2W_{w/o\ G}$ | 12.28 | 1.15 | 0.81 | 0.99 | 0.76 | | $W2W_{w/o\ G}$ | 11.88 | 2.84 | 3.62 | 2.68 | 2.50 |
| stove | W2W | 16.15 | 2.63 | 2.64 | 1.94 | 2.72 | telephone | W2W | 14.09 | 1.18 | 0.96 | 1.05 | 0.96 |
| | $W2W_{w/o\ G}$ | 16.13 | 3.06 | 4.30 | 3.08 | 2.98 | | $W2W_{w/o\ G}$ | 13.42 | 1.17 | 1.50 | 1.46 | 1.38 |
| steep | W2W | 5.89 | 0.63 | 0.39 | 0.53 | 0.42 | bamboo | W2W | 14.54 | 2.02 | 1.20 | 0.76 | 1.02 |
| | $W2W_{w/o\ G}$ | 7.30 | 1.46 | 2.42 | 0.87 | 1.93 | | $W2W_{w/o\ G}$ | 15.40 | 3.01 | 1.38 | 1.09 | 1.42 |
| warm | W2W | 7.79 | 0.68 | 0.69 | 0.68 | 0.69 | brush | W2W | 11.17 | 1.88 | 2.13 | 1.81 | 2.45 |
| | $W2W_{w/o\ G}$ | 8.67 | 1.05 | 1.01 | 0.79 | 0.85 | | $W2W_{w/o\ G}$ | 13.69 | 2.51 | 2.89 | 2.39 | 2.83 |
| aged | W2W | 13.72 | 0.50 | 0.53 | 0.39 | 0.66 | button | W2W | 4.73 | 2.37 | 2.08 | 1.82 | 2.01 |
| | $W2W_{w/o\ G}$ | 13.50 | 0.77 | 0.94 | 0.78 | 0.93 | | $W2W_{w/o\ G}$ | 5.94 | 3.25 | 3.19 | 2.54 | 2.74 |
| pizza | W2W | 10.70 | 1.47 | 1.07 | 1.19 | 0.90 | | | | | | | |
| | $W2W_{w/o\ G}$ | 9.59 | 2.21 | 2.54 | 1.25 | 1.18 | | | | | | | |

表7:1クラス漸増学習における未見語のlog G-PPL (All-Protocol)、各未見語のサンプルサイズは8から32の範囲。

| # Samples | | 0 | 8 | 16 | 24 | 32 | # Samples | | 0 | 8 | 16 | 24 | 32 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| crosswalk | W2W | 10.82 | 8.48 | 7.43 | 7.70 | 5.95 | donkey | W2W | 8.70 | 0.84 | 0.81 | 0.67 | 0.79 |
| | $W2W_{w/o\ G}$ | 10.91 | 10.88 | 7.53 | 7.15 | 7.5 | | $W2W_{w/o\ G}$ | 9.69 | 1.97 | 1.99 | 2.35 | 2.01 |
| cheese | W2W | 12.16 | 2.62 | 3.00 | 1.27 | 1.04 | barefoot | W2W | 9.71 | 6.93 | 4.58 | 5.55 | 6.27 |
| | $W2W_{w/o\ G}$ | 13.07 | 2.81 | 3.13 | 2.56 | 1.49 | | $W2W_{w/o\ G}$ | 9.95 | 6.52 | 4.67 | 5.74 | 5.88 |
| star | W2W | 8.70 | 1.49 | 1.47 | 1.09 | 1.18 | elephant | W2W | 15.24 | 1.44 | 1.65 | 1.81 | 1.44 |
| | $W2W_{w/o\ G}$ | 10.59 | 2.93 | 2.10 | 1.99 | 1.39 | | $W2W_{w/o\ G}$ | 14.75 | 2.17 | 1.98 | 1.73 | 1.61 |
| classroom | W2W | 3.96 | 0.47 | 0.36 | 0.43 | 0.32 | heart | W2W | 9.34 | 2.97 | 1.90 | 1.76 | 1.76 |
| | $W2W_{w/o\ G}$ | 5.10 | 0.95 | 0.88 | 1.05 | 0.95 | | $W2W_{w/o\ G}$ | 9.31 | 2.99 | 2.50 | 2.65 | 2.96 |
| fluffy | W2W | 16.44 | 1.88 | 1.78 | 0.82 | 1.36 | gym | W2W | 5.13 | 2.14 | 0.44 | 0.74 | 0.69 |
| | $W2W_{w/o\ G}$ | 15.61 | 1.83 | 1.71 | 1.37 | 1.47 | | $W2W_{w/o\ G}$ | 4.88 | 3.73 | 1.30 | 1.08 | 1.45 |
| circular | W2W | 15.21 | 1.59 | 1.07 | 1.55 | 1.23 | security | W2W | 15.08 | 1.07 | 0.81 | 1.28 | 0.71 |
| | $W2W_{w/o\ G}$ | 15.12 | 2.25 | 2.25 | 1.81 | 1.61 | | $W2W_{w/o\ G}$ | 14.75 | 1.50 | 1.22 | 1.53 | 1.17 |
| sink | W2W | 14.23 | 1.17 | 0.92 | 1.11 | 1.38 | cafe | W2W | 6.28 | 1.90 | 1.38 | 1.98 | 1.39 |
| | $W2W_{w/o\ G}$ | 15.49 | 1.84 | 1.65 | 1.60 | 1.84 | | $W2W_{w/o\ G}$ | 7.03 | 2.17 | 1.92 | 2.08 | 1.72 |
| doctor | W2W | 13.03 | 1.17 | 1.05 | 1.38 | 1.18 | teacher | W2W | 16.68 | 1.95 | 2.15 | 1.52 | 1.48 |
| | $W2W_{w/o\ G}$ | 12.44 | 1.17 | 1.23 | 1.39 | 1.58 | | $W2W_{w/o\ G}$ | 16.08 | 2.68 | 2.37 | 1.85 | 1.83 |
| foreign | W2W | 9.48 | 0.62 | 0.95 | 0.85 | 0.47 | student | W2W | 16.28 | 1.38 | 1.07 | 1.20 | 1.03 |
| | $W2W_{w/o\ G}$ | 10.01 | 1.03 | 0.88 | 1.18 | 0.95 | | $W2W_{w/o\ G}$ | 16.52 | 2.21 | 1.29 | 1.40 | 1.61 |
| diverse | W2W | 16.44 | 0.60 | 0.22 | 0.52 | 0.24 | newborn | W2W | 16.43 | 1.71 | 0.88 | 0.91 | 1.11 |
| | $W2W_{w/o\ G}$ | 16.05 | 0.81 | 0.65 | 0.97 | 0.65 | | $W2W_{w/o\ G}$ | 16.30 | 2.02 | 1.32 | 1.61 | 1.76 |
| product | W2W | 10.25 | 0.84 | 0.75 | 1.39 | 1.15 | pan | W2W | 12.04 | 1.70 | 2.12 | 1.87 | 2.02 |
| | $W2W_{w/o\ G}$ | 12.28 | 1.15 | 0.81 | 0.99 | 0.76 | | $W2W_{w/o\ G}$ | 11.88 | 2.84 | 3.62 | 2.68 | 2.50 |
| stove | W2W | 16.15 | 2.63 | 2.64 | 1.94 | 2.72 | telephone | W2W | 14.09 | 1.18 | 0.96 | 1.05 | 0.96 |
| | $W2W_{w/o\ G}$ | 16.13 | 3.06 | 4.30 | 3.08 | 2.98 | | $W2W_{w/o\ G}$ | 13.42 | 1.17 | 1.50 | 1.46 | 1.38 |
| steep | W2W | 5.89 | 0.63 | 0.39 | 0.53 | 0.42 | bamboo | W2W | 14.54 | 2.02 | 1.20 | 0.76 | 1.02 |
| | $W2W_{w/o\ G}$ | 7.30 | 1.46 | 2.42 | 0.87 | 1.93 | | $W2W_{w/o\ G}$ | 15.40 | 3.01 | 1.38 | 1.09 | 1.42 |
| warm | W2W | 7.79 | 0.68 | 0.69 | 0.68 | 0.69 | brush | W2W | 11.17 | 1.88 | 2.13 | 1.81 | 2.45 |
| | $W2W_{w/o\ G}$ | 8.67 | 1.05 | 1.01 | 0.79 | 0.85 | | $W2W_{w/o\ G}$ | 13.69 | 2.51 | 2.89 | 2.39 | 2.83 |
| aged | W2W | 13.72 | 0.50 | 0.53 | 0.39 | 0.66 | button | W2W | 4.73 | 2.37 | 2.08 | 1.82 | 2.01 |
| | $W2W_{w/o\ G}$ | 13.50 | 0.77 | 0.94 | 0.78 | 0.93 | | $W2W_{w/o\ G}$ | 5.94 | 3.25 | 3.19 | 2.54 | 2.74 |
| pizza | W2W | 10.70 | 1.47 | 1.07 | 1.19 | 0.90 | | | | | | | |
| | $W2W_{w/o\ G}$ | 9.59 | 2.21 | 2.54 | 1.25 | 1.18 | | | | | | | |

Table 7: The log G-PPL (All-Protocol) of unseen words in one-class incremental learning, each unseen word with a sample size ranging from 8 to 32.

**A For every submission:**

3 A1. 仕事の限界について説明しましたか?セクション7、制限事項

7 A2. あなたの研究の潜在的なリスクについて議論しましたか?この研究には、人間の主題や人間の研究は含まれていません。本研究では、問題設定と計算フレームワークを提案しており、当面は実世界のアプリケーションに展開することはできない。

3 A3. 抄録と紹介は、本論文の主な主張をまとめたものであるか。第1章

7 A4. この論文に取り組む際に、AIライティングアシスタントを使用しましたか?空白は左。

B 3 科学的成果物を使用または作成したか?

*Section 2*

3 B1. 使用した成果物の作成者を引用しましたか?セクション 2.4

3 B2. 3 B2. プラグインや利用規約、/または配布に関する説明がありましたか?コードリリースと一緒に掲載されます。

3 B3. 既存の成果物の使用が、意図した使用と一致しているかどうか、指定されていれば、議論しましたか?作成した成果物について、意図した使用方法と、それが元のアクセス条件と適合しているかどうかを明記していますか(特に、研究目的でアクセスしたデータの派生物は、研究コンテキスト以外で使用しないでください)?コードリリースと一緒に含める予定です。

□ B4. 収集/使用されたデータに、個人または攻撃的なコンテンツの名称または一意に識別する情報が含まれているかどうか、また、保護/匿名化のために取られた手順について説明しましたか?該当なし。左空欄。

□ B5. ドメイン、言語、言語現象、代表的な人口動態グループなどを網羅した成果物の記録を提供 しましたか?該当なし。左空欄。

3 B6. 使用したデータについて、例数、train / test / dev split の詳細など、関連する統計情報を報告しましたか?一般的に使用されているベンチマークデータセットでも、訓練/検証/テスト分割の例数は、読者が実験結果を理解するために必要な文脈を提供するため、記載しています。例えば、大規模なテストセットでは精度の差が小さくても、小規模なテストセットではそうならないことがあります。セクション 2 と付録 A

C 3 計算機実験をしましたか?

*Section 4*

3 C1. 使用したモデルのパラメータ数、総計算予算(例:GPU 時間)、使用したコンピューティングインフラを報告しましたか?付録C

ACL 2023 で使用された Responsible NLP Checklist は、
*assistance.*

## ACL 2023 Responsible NLP Checklist

### A For every submission:

☑ A1. Did you describe the limitations of your work?
*Section 7, Limitations*

☒ A2. Did you discuss any potential risks of your work?
*This study does not contain any human subjects or human studies. The study proposes a problem formulation and a computational framework, which is not deployable to any real-world applications in the foreseeable future.*

☑ A3. Do the abstract and introduction summarize the paper's main claims?
*Section 1*

☒ A4. Have you used AI writing assistants when working on this paper?
*Left blank.*

### B ☑ Did you use or create scientific artifacts?

*Section 2*

☑ B1. Did you cite the creators of artifacts you used?
*Section 2.4*

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
*Will be included along with the code release*

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
*Will be included along with the code release*

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
*Not applicable. Left blank.*

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
*Not applicable. Left blank.*

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
*Section 2 and Appendix A*

### C ☑ Did you run computational experiments?

*Section 4*

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
*Appendix C*

*The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.*

AI writing 3 C2 に関する質問を追加し、NAACL 2022 から採用したものである。ハイパーパラメータ探索と最適なハイパーパラメータ値を含む実験セットアップについて議論しましたか?付録C

7 C3. 結果に関する記述統計(例:結果に関するエラーバー、実験セットからの要約統計)を報告しましたか、また、最大値、平均値などを報告しているか、それとも1回の実行で報告しているかは明らかですか?この研究では、繰り返し実行しても経済的に実行不可能な事前学習フレームワークを含んでいます。

3 C4. 既存のパッケージ(前処理、正規化、評価など)を使用した場合、使用した実装、モデル、パラメータ設定(NLTK、Spacy、ROUGEなど)を報告しましたか?付録C

**D** 7 人間のアノテーター(クラウドワーカーなど)や研究を人間の参加者に利用しましたか?

*Left blank.*

□ D1. スクリーンショット、参加者やアノテーターへのリスクの否認など、参加者に与えられた指示の全文を報告しましたか?該当なし。左空欄。

□ D2. クラウドソーシングプラットフォーム、学生など、どのように募集し、有料化した参加者について情報を報告し、参加者の人口統計学(居住国など)を考慮して、その支払いが適切であるかどうかを議論しましたか?該当なし。該当なし。左空欄。

□ D3. D3. データを利用する/利用する人から同意を得たかどうか、またどのように得たかについて議論しましたか?例えば、クラウドソーシングでデータを収集した場合、クラウドワーカーへの指示でデータがどのように使用されるかを説明したか?該当なし。空白は除く。

□ D4. データ収集プロトコルは倫理審査委員会により承認(または免除と判断)されたか。該当なし。空白。

□ D5. データの出所となるアノテーター集団の基本的な人口統計学的および地理的特性を報告しましたか?該当なし。左空欄。

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*Appendix C*

☒ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
*The study involves a pre-training framework which is not economically feasible for repeated runs.*

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
*Appendix C*

**D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?**

*Left blank.*

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
*Not applicable. Left blank.*

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
*Not applicable. Left blank.*

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
*Not applicable. Left blank.*

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*Not applicable. Left blank.*

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
*Not applicable. Left blank.*