CLIP と不均衡最適輸送を用いた画像–文章間の類似性評価および局所アライメントの可視化

志子田直輝　五藤巧　西田悠人　坂井優介　上垣外英剛　渡辺太郎
奈良先端科学技術大学院大学
{shikoda.naoki.sm1, goto.takumi.gv7, nishida.yuto.nu8}@naist.ac.jp
{sakai.yusuke.sr9, kamigaito.h, taro}@is.naist.jp

概要

CLIP は，画像と文章を同一の空間へ写像して学習することで共通の埋め込み表現を獲得し，表現間のコサイン類似度に基づく CLIPScore による画像–文章間の大局的な類似性評価を可能とした．一方で，この表現は入力の局所領域間の関係を捉えられていないことが指摘されている．本研究では，CLIP および派生モデルの出力に基づく新たな画像–文章間の類似性評価，および局所領域間のアライメントの可視化手法を提案する．評価実験により，提案法は CLIPScore などのベースラインと同等以上の人手評価との相関を示し，さらに局所対応を可視化できることを確認した．

1　はじめに

画像と自然言語の対応関係を学習した視覚言語モデルは，マルチモーダルな基盤モデルとして幅広い下流タスクに応用されている．特に視覚言語モデル内で画像と文章とを対応付ける手法として，入力を同じ埋め込み空間に写像するというシンプルな構造の CLIP [1] がよく用いられる．下流タスクでは一般に，CLIP が画像と文章それぞれの全体表現を対照学習して得た表現が利用される．

しかし，モデルが特別に設計された構造でない限り，直接 CLIP から局所的な表現を得ることはできない．また，単一のベクトルとして得られる全体表現には局所的な情報が反映されていないおそれがある．既存研究では，CLIP の出力を元にして画像–文章間の類似度を測る CLIPScore [2] が，画像の局所的な構成や関係性を扱えないこと [3]，文や画像の構造を捉えられていないこと [4] が指摘されている．もしも，全体表現だけでなく画像–文章間の局所対応を考慮したスコアリングを行うことができれば，これらの問題を解決する糸口になると考えられる．

そこで本研究では，事前学習済みの CLIP およびその派生モデルから得られる表現から，最適輸送を用いて画像–文章間の局所アライメントを推定し，それに基づく画像–文章間の類似度評価，およびアライメントの可視化手法を提案する．類似度評価で計算したスコアと人手評価との相関はベースラインと同等以上となり，得られたアライメントの可視化を通して，これらのモデルがどのように局所対応を捉えているのかを確認した．

2　背景

2.1　CLIPScore

CLIPScore [2] は，CLIP [1] を用いて，画像–文章間の関連度を参照文を用いずに測る評価手法である．他の評価指標として，参照文との n-gram を利用する METEOR [5], CIDEr [6] や，シーングラフを利用する SPICE [7] などが存在した．これらに比べて CLIPScore は，参照文が不要ながら人手評価との高い相関を示す．CLIPScore は式 (1) で表される．ここで，(I, T) を入力画像–文章のペアとする．また，L2 正規化を $\mathrm{Norm}_{\mathrm{L2}}(\boldsymbol{x}) = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2}$, コサイン類似度を $\mathrm{cossim}(\boldsymbol{x}, \boldsymbol{y})$ と定義する．

$$\begin{aligned}&\mathrm{CLIPScore}(I, T)\\ &\quad = w * \langle \mathrm{Norm}_{\mathrm{L2}}(\mathrm{Enc}_{\mathrm{i}}(I)), \mathrm{Norm}_{\mathrm{L2}}(\mathrm{Enc}_{\mathrm{t}}(T)) \rangle \\ &\quad = w * \mathrm{cossim}(\mathrm{Enc}_{\mathrm{i}}(I), \mathrm{Enc}_{\mathrm{t}}(T))\end{aligned} \tag{1}$$

$\mathrm{Enc}_{\mathrm{i}}, \mathrm{Enc}_{\mathrm{t}}$ はそれぞれ CLIP の画像側, 文章側のエンコーダを表し，$\mathrm{Enc}_{\mathrm{i}}(I), \mathrm{Enc}_{\mathrm{t}}(T)$ は，それぞれの入力から得た埋め込み表現である．一般に，値をおおよそ 0 から 1 にスケーリングする係数 w として $w = 2.5$ が利用されている．

2.2 CLIP の課題

CLIP は，画像と文章をそれぞれ独立したエンコーダによって同一の空間へ写像し，大規模なデータを用いて正しい画像–文章ペアを正例，それ以外を負例とした対照学習を行う．また，CLIPScore により画像–文章間の大局的な類似性を評価できる．

しかし，CLIPScore が画像の局所的な構成性を扱えないこと [3]，構造を十分に捉えられていないこと [4] が指摘されている．CLIP の出力は大局的な表現にすぎず，そこから画像中の領域や文章中の単語といった局所情報が表現や類似度へ与える影響，および局所間の対応関係を解析することは容易ではない．CLIP が注目した局所情報の可視化にはいくつかの手法 [8, 9, 10, 11] が挙げられるが，手法によっては追加の学習が必要，モデルの中間層を取得するための固有の実装が必要，エンコーダの注目した情報は得られるが画像–文章間のアライメントは得られない，などの課題がある．ここで，(1) 追加学習が不要であること (2) 設計がモデルの内部構造に依存しないこと (3) 画像と文章の局所間の関係が抽出できること という 3 点を満たす手法が求められる．

2.3 最適輸送

最適輸送（Optimal Transport; OT）[12] は，2 つの分布を最適輸送問題として比較する手法である．通常の OT では分布間を過不足なく輸送するよう最適化されるが，不均衡最適輸送（Unbalanced Optimal Transport; UOT）[13] では過不足を許容した最適化が可能となる．2 つの分布 $\mathbf{a}, \mathbf{b} \quad (\forall a_i, b_j \in \mathbb{R}_{\geq 0})$ において，その間の輸送コストを $\mathbf{C} \in \mathbb{R}^{|\mathbf{a}|\times|\mathbf{b}|}$ とすると，輸送計画 $\mathbf{P} \in \mathbb{R}_{\geq 0}^{|\mathbf{a}|\times|\mathbf{b}|}$ は式 (2) のように定式化される．

$$\begin{aligned}\mathbf{P} &= \mathrm{UOT}(\mathbf{a}, \mathbf{b}, \mathbf{C}) \\ &= \underset{\mathbf{P}\in\mathbb{R}_{\geq 0}^{|\mathbf{a}|\times|\mathbf{b}|}}{\arg\min} \sum_{i=1}^{|\mathbf{a}|}\sum_{j=1}^{|\mathbf{b}|} P_{i,j}\cdot C_{i,j} + \mathrm{reg}\cdot \mathrm{KL}(\mathbf{P}, \mathbf{a}\mathbf{b}^\top) \\ &\quad + \mathrm{reg}_{\mathrm{m1}}\cdot \mathrm{KL}(\mathbf{P}\mathbb{1}_{|\mathbf{a}|}, \mathbf{a}) + \mathrm{reg}_{\mathrm{m2}}\cdot \mathrm{KL}(\mathbf{P}^\top\mathbb{1}_{|\mathbf{b}|}, \mathbf{b})\end{aligned} \tag{2}$$

ここで，$\mathrm{KL}(\cdot,\cdot)$ は分布間の Kullback-Leibler divergence であり，reg は均一的な輸送を強制する度合いを調整する係数，$\mathrm{reg}_{\mathrm{m1}}, \mathrm{reg}_{\mathrm{m2}}$ は輸送時の過不足の許容量を調整する係数である．得られる輸送計画の各要素 $P_{i,j}$ は a_i, b_j 間の対応の度合いを表す．

本研究では画像–文章間の領域の対応を，それぞれの局所情報の輸送として捉える．これに最適輸送を用いることで，局所的な対応を得ることが期待される．ただし，画像–文章間の全ての領域が互いに対応するとは限らず，文章が画像の前景など一部分のみを記述している，あるいは画像に含まれない内容を記述していることがあるため，情報の過不足を許容する UOT を用いた．

3 提案法

2.2 節で述べたように，(1) 追加学習が不要であること (2) 設計がモデルの内部構造に依存しないこと (3) 画像と文章の局所間の関係が抽出できること，という 3 点を満たす手法を考える．本研究では (1) 既存の事前学習済みモデルをそのまま利用し，(2) Leave-one-out [14] をもとに出力の観察のみから局所情報を推定し，(3) 最適輸送によって局所間のアライメントを計算することで，上記 3 点を満たす新たな評価手法を提案する．

3.1 Leave-one-out に基づく設計

あるブラックボックスなモデルにおいて，入力の局所情報が出力へ与える影響を推定したい．そこで，トークンを 1 つ削除した際の出力の変化からそのトークンの重要度を推定できる Leave-one-out をもとに，モデルへの入力 I, T とそれぞれの入力におけるトークン削除後の入力 $I_{\setminus(i,j)}, T_{\setminus k}$ との出力の差分を局所情報として推定し，利用した．

文章側では，各単語を独立に削除したものをトークン削除後の入力とする．単語数 L の入力文 $T = (t_1, \ldots, t_L)$ に対して，k 番目の単語を削除した後の単語列全体 $T_{\setminus k}$ を式 (3) のように定式化する．

$$T_{\setminus k} = (t_1, \ldots, t_{k-1}, t_{k+1}, \ldots, t_L) \tag{3}$$

例えば，入力文 T が (a, dog, rolling, in, the, grass, .) のとき，3 番目の単語の削除後の単語列は $T_{\setminus 3}$ = (a, dog, in, the, grass, .) で表す．

CLIP およびその派生モデルにおいて，画像側のエンコーダとして利用される Vision Transformer [15] では，入力画像をパッチ単位で捉え，それらをトークン列のようにして扱う．そこで画像側では，画像 I のパッチ $p_{(i,j)}$ を独立にガウシアンノイズ p' へ置換することで，そのパッチの削除とみなした．入力画像 $I = (p_{(1,1)}, \ldots, p_{(N_\mathrm{h}, N_\mathrm{w})})$ に対するパッチ削除後の画像全体 $I_{\setminus(i,j)}$ を式 (4) のように定義する．

$$I_{\setminus(i,j)} = (p_{(1,1)}, \ldots p_{(i,j-1)}, p', p_{(i,j+1)}, \ldots, p_{(N_\mathrm{h}, N_\mathrm{w})}) \tag{4}$$

ここで，$N_\mathrm{h}, N_\mathrm{w}$ はパッチの分割数である．

以上の操作により，文章側・画像側のどちらにおいても，それぞれ削除したトークンが出力に与えていた影響を推定した．

3.2 UOT による対応付け

UOT のための $\mathbf{a}, \mathbf{b}, \mathbf{C}$ を，次のように定義する．

$$a_k = \max(0, \mathrm{s}(I,T) - \mathrm{s}(I,T_{\setminus k})) \tag{5}$$

$$b_{(i,j)} = \max(0, \mathrm{s}(I,T) - \mathrm{s}(I_{\setminus(i,j)},T)) \tag{6}$$

$$C_{k,(i,j)} = 1 - \mathrm{cossim}(\mathrm{Enc}_t(T_{\setminus k}), \mathrm{Enc}_i(I_{\setminus(i,j)})) \tag{7}$$

$$(k = 1,\ldots,L), \quad ((i,j) = (1,1),\ldots,(N_\mathrm{h},N_\mathrm{w}))$$

ここで $s(I,T)$ は CLIPScore(I,T) など，画像 I-文章 T ペアに何らかのスコアを返す関数，L は文章の単語長, N は画像のパッチ数である．Leave-one-out に基づき $I_{\setminus(i,j)}, T_{\setminus k}$ を求め，$\mathbf{a}$, $\mathbf{b}$, $\mathbf{C}$ を計算した．$a_k, b_{(i,j)}$ は式 (5,6) で定め，それぞれのトークンを削除したことによるのスコアの減少値とした．輸送コスト $C_{k,(i,j)}$ は式 (7) で定め，$T_{\setminus k}, I_{\setminus(i,j)}$ 間のコサイン類似度を 1 から引いたものとした．

3.3 輸送結果を用いた類似度評価

得られる輸送計画は，単語とパッチの対応付けを表すことが期待される．そこで，式 (5, 6, 7) により定義した $\mathbf{a}, \mathbf{b}, \mathbf{C}$ で不均衡最適輸送を行い，輸送計画 $\mathbf{P} = \mathrm{UOT}(\mathbf{a}, \mathbf{b}, \mathbf{C})$ から画像–文章間の類似度を表すスコアを求めた．スコアの算出は五藤ら [16] が提案した手法に従い，$\mathbf{a}, \mathbf{b}$ および $\mathbf{P}$ を用いて式 (8) で求められる F 値とした．なお，本稿では $\beta = 0.5$ を用いた．UOT のハイパーパラメータである reg および $\mathrm{reg}_{\mathrm{m1}}, \mathrm{reg}_{\mathrm{m2}}$ は，モデルごとに Optuna [17] による探索を行った．詳細は Appendix A に示す．

$$\begin{aligned}
&\mathrm{TP} = \sum_{i,j} P_{i,j}, \ \mathrm{FP} = \sum_i a_i - \mathrm{TP}, \ \mathrm{FN} = \sum_j b_j - \mathrm{TP}, \\
&\mathrm{Prec} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \quad \mathrm{Rec} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \\
&\mathrm{UOTScore}(\mathbf{P}) = F_\beta = \frac{(1+\beta^2)\cdot \mathrm{Prec}\cdot \mathrm{Rec}}{\beta^2 \cdot \mathrm{Prec} + \mathrm{Rec}}.
\end{aligned} \tag{8}$$

4 実験

4.1 モデル

実装には OpenCLIP [18] を利用した．ベースラインには LAION-5B [19] で事前学習された CLIP の他，派生モデルとして SigLIP [20] とその改良版である SigLIP2 [21] を用いた．画像側エンコーダの構造として，CLIP は ViT-B/16[1] および ViT-B/32[2] を，SigLIP と SigLIP2 は ViT-B/16[3][4] を使用した．CLIP でのスコア関数は CLIPScore を用いた．同様の評価を行うため，SigLIP および SigLIP2 では推論時の実装に基づきスコア関数を式 (9) で定義した．

(a) A young person kisses an old person. **(b)** An old person kisses a young person.

図 1: Winoground の画像ペアの例

$$\begin{aligned}
&\mathrm{SigLIPScore}(I,T) \\
&= \sigma(\langle \mathrm{Norm}_{\mathrm{L2}}(\mathrm{Enc}_\mathrm{i}(I)), \mathrm{Norm}_{\mathrm{L2}}(\mathrm{Enc}_\mathrm{t}(T))\rangle),
\end{aligned} \tag{9}$$

ここで $\sigma(\cdot)$ は Sigmoid 関数である．これを便宜上 SigLIPScore, SigLIP2Score と呼ぶ．

4.2 データセットと評価指標

人間によるスコアが付与されたデータセットとして Flickr8k-Expert と Flickr8k-CF [22] を，構造が複雑な画像–文章データセットとして Winoground [4] を用いた．スコアリングの評価として，Flickr8k では人手評価との相関を測った．Winoground では，Winoground 内で定義された指標に基づき評価した．

各データセットの詳細を以下に述べる．

Flickr8k-Expert 専門家 3 名が各画像–文章ペアを 4 段階で評価したデータセットである．本稿では 3 名の平均点を人間のスコアとし，画像 1,000 枚，説明文 977 件からなる計 5,822 ペアを使用した．

Flickr8k-CF クラウドワーカーが各画像–文章ペアについて，「キャプションが画像を妥当に記述しているか」を “Yes” と “No” の二値で判断したデータセットである．本稿では各ペアごとの “Yes” の割合を人間のスコアとし，画像 1,000 枚，説明文 1,000 件からなる計 47,380 ペアを使用した．

Winoground Winoground [4] は，図 1 に示すような，物体の構成を入れ替えた 2 つの画像ペア (I_0, I_1)

1) https://huggingface.co/laion/CLIP-ViT-B-16-laion2B-s34B-b88K
2) https://huggingface.co/laion/CLIP-ViT-B-32-laion2B-s34B-b79K
3) https://huggingface.co/timm/ViT-B-16-SigLIP
4) https://huggingface.co/timm/ViT-B-16-SigLIP2

表 1: 評価対象の Flickr8k の kendall τ_b の値

Model	Flickr8k-Ex	Flickr8k-CF
CLIPScore ViT-B/32	0.584	0.345
→ UOTScore (ours)	**0.605**	**0.374**
CLIPScore ViT-B/16	0.599	0.350
→ UOTScore (ours)	**0.617**	**0.388**
SigLIPScore ViT-B/16	0.603	0.362
→ UOTScore (ours)	**0.629**	**0.388**
SigLIP2Score ViT-B/16	**0.607**	**0.369**
→ UOTScore (ours)	0.594	0.367

と，対応して文章の語を入れ替えた意味の異なる 2 つの説明文 (T_0, T_1) からなる．正しいペアの選択には画像と文章のどちらにも構造把握が要求されるデータセットである．本稿では，画像 800 枚・説明文 800 件からなる計 1,600 ペアを使用した．

4.3 人手評価との相関

人手評価との相関を測る指標として kendall の順位相関係数 (τ_b) を用いた．Flickr8k-Expert, Flickr8k-CF の各画像–文章ペアに対して，ベースラインおよび提案法によるスコアを計算し，人手評価との順位相関を測った．結果を表 1 に示す．

Flickr8k の人手評価との相関は，SigLIP2 を除き，ベースラインとなる元の CLIPScore や SigLIPScore を上回った．さらに，得られた輸送計画が局所情報の対応を表すことで，モデルが捉えた画像–文章間の局所間の関係性を可視化できるようになった．CLIP ViT-B/16 のスコアリング時に得られたアライメントを図 2, 3 に示す．ここで，図の垂直方向は単語，水平面は画像であり，画像のパッチに輸送された量を単語ごとに色分けし散布図として可視化している．球が大きいほどその単語-画像パッチ間の関連性が大きいと考えられる．

図 2, 3 を見ると，画像に対して CLIPScore の高い文では，“grass” と芝生との対応が大きいこと，“dog” は対応先がない，あるいは犬周辺の芝と対応付いていることがわかる．CLIPScore の低い文では，関連する単語がない場合は画像全体に広く対応付くこと，画像に合わない文章でも “bare (feet)” のような単語では犬の足元部分が対応していることがわかる．

これらはベースラインの変動値をもとにしたアライメントであり，モデルが入力の対応関係をどのように捉えているかを可視化しているといえる．

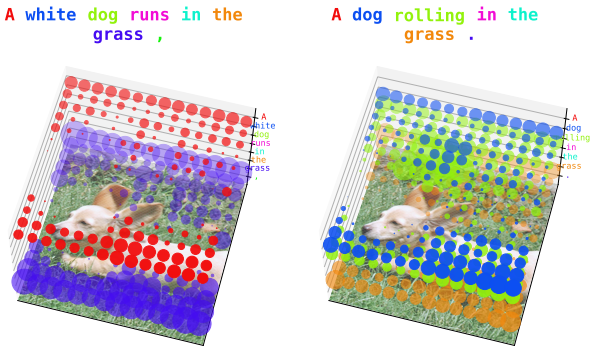

図 2: CLIPScore の高いペアの可視化結果

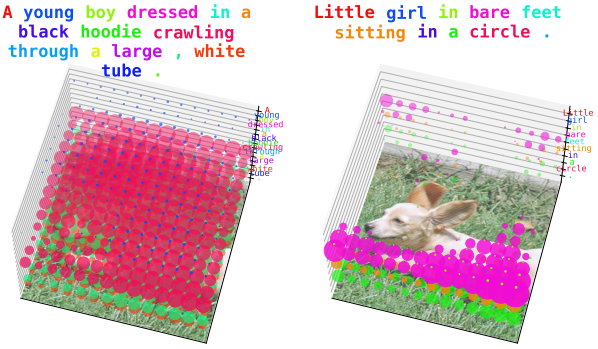

図 3: CLIPScore の低いペアの可視化結果

4.4 Winoground による評価

得られた画像–文章間の局所情報の対応付けを確認するため，Winoground による各指標の評価および輸送結果の可視化を行った．詳細を Appendix B に示す．Flickr8k の結果と異なり，Winoground の各指標はベースラインと同等程度の結果となった．

5 おわりに

CLIP および派生モデルの局所情報の分析を行うため，Leave-one-out による出力の差分で局所情報を推定し，最適輸送を用いた局所アライメントの算出と類似度スコアリングを行った．提案法によるスコアリングは，人手評価との相関でベースラインと同等以上の結果となり，最適輸送がこれらのモデルに対して適用可能であること，また事前学習済みモデルのスコアリングを，追加の学習をせずに改善できることを示した．得られたアライメントをモデルの捉えた局所領域間の対応とみることで，モデルがどのように画像–文章間の対応を獲得しているかを可視化した．複雑な構造を持つ Winoground による評価ではベースラインと同程度であり，可視化結果から CLIP の持つ課題が示唆された．

参考文献

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, **Proceedings of the 38th International Conference on Machine Learning**, Vol. 139 of **Proceedings of Machine Learning Research**, pp. 8748–8763. PMLR, 18–24 Jul 2021.

[2] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 7514–7528, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[3] Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. An explainable toolbox for evaluating pre-trained vision-language models. In **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 30–37, 2022.

[4] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 5238–5248, June 2022.

[5] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss, editors, **Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization**, pp. 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[6] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation . In **2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 4566–4575, Los Alamitos, CA, USA, June 2015. IEEE Computer Society.

[7] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, **Computer Vision – ECCV 2016**, pp. 382–398, Cham, 2016. Springer International Publishing.

[8] Boris Joukovsky, Fawaz Sammani, and Nikos Deligiannis. Model-agnostic visual explanations via approximate bilinear models. In **2023 IEEE International Conference on Image Processing (ICIP)**, pp. 1770–1774, 2023.

[9] Yi Li, Hualiang Wang, Yiqun Duan, Jiheng Zhang, and Xiaomeng Li. A closer look at the explainability of contrastive language-image pre-training. **Pattern Recognition**, Vol. 162, p. 111409, 2025.

[10] Chenyang Zhao, Kun Wang, Xingyu Zeng, Rui Zhao, and B. Antoni Chan. Gradient-based visual explanation for clip. In **International Conference on Machine Learning (ICML)**, July 2024.

[11] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In **Proceedings of the British Machine Vision Conference (BMVC)**, 2018.

[12] Leonid V Kantorovich. On the translocation of masses. **Manage. Sci.**, Vol. 5, No. 1, p. 1–4, October 1958.

[13] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a wasserstein loss. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 28. Curran Associates, Inc., 2015.

[14] Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. Pathologies of neural models make interpretations difficult. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 3719–3728, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. **ICLR**, 2021.

[16] 五藤巧, 坂井優介, 渡辺太郎. 文法誤り訂正における編集ベクトルの最適輸送に基づく性能評価尺度. 言語処理学会 第32回年次大会, 2026.

[17] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In **Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining**, KDD '19, p. 2623–2631, New York, NY, USA, 2019. Association for Computing Machinery.

[18] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021.

[19] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In **Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track**, 2022.

[20] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In **Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)**, pp. 11975–11986, October 2023.

[21] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025.

[22] P. Young M. Hodosh and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. In **Journal of Artifical Intellegence Research (JAIR)**, pp. 853–899, 2013.

表 2: 探索したペア数と評価対象数

Dataset	探索数	評価対象数	評価対象外数
Flickr8k-Expert	291	2,970	2,852
Flickr8k-CF	0	26,808	21,022
Winoground	0	1,600	0

表 3: 探索で得られた UOT のハイパーパラメータ

BaseModel	reg	reg_{m1}	reg_{m2}
CLIP ViT-B/32	1.693e-02	1.752e-09	3.287e-09
CLIP ViT-B/16	1.703e-02	2.464e-04	3.145e-11
SigLIP ViT-B/16	1.976e-02	1.906e-07	7.311e-12
SigLIP2 ViT-B/16	4.071e-01	2.575e-09	1.758e-09

A Optuna によるパラメータ探索

Optuna で UOT のハイパーパラメータを探索した．探索データとして，Flickr8k-Expert に含まれるペアのうち 5% を使用した．探索したデータと評価するデータとの重複を避けるため，残りの Flickr8k データセットのうち，探索データと同じ画像や文章が含まれるペアは対象外として，ベースラインおよび UOTScore の評価を行った．詳細を表 2 に示す．また，探索によって得られたハイパーパラメータを表 3 に示す．

B Winoground による分析

Winoground における各指標 Winoground では，視覚言語モデルが画像，文章ごとに正しいペアを選択したか評価する指標として TextScore, ImageScore, GroupScore を式 (10, 11, 12) で定義している．

$$\text{TextS}(I_0, T_0, I_1, T_1) = \begin{cases} 1 & \text{if } s(I_0, T_0) > s(I_0, T_1) \\ & \quad \text{and } s(I_1, T_1) > s(I_1, T_0) \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

$$\text{ImageS}(I_0, T_0, I_1, T_1) = \begin{cases} 1 & \text{if } s(I_0, T_0) > s(I_1, T_0) \\ & \quad \text{and } s(I_1, T_1) > s(I_0, T_1) \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

$$\text{GroupS}(I_0, T_0, I_1, T_1) = \begin{cases} 1 & \text{if TextS}(I_0, T_0, I_1, T_1) \\ & \quad \text{and ImageS}(I_0, T_0, I_1, T_1) \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

性能評価 ベースライン，提案法における各指標の結果を表 4 に示す．人手評価にあたる MTurk Human の値は Winoground [4] の論文値を引用した．指標上は Flickr8k と異なり，ベースラインと同程度の結果となった．

アライメントの可視化 図 4, 5 に，Winoground のペアを入力して得られたアライメントを示す．“young”, “old” に着目すると，正しいペアではそれぞれ続く “person” と同じパッチに対する球がみられるのに対して，誤ったペアでは特に “young”, “old” の球が小さくなり，さらに一方は “person” の球を失っているなど，局所領域をうまく対応付けられていないようにみえる．

表 4: TextScore, ImageScore, GroupScore の値

Model	Text	Image	Group
MTurk Human	**89.50**	**88.50**	**85.50**
Random Chance	25.00	25.00	16.67
CLIPScore ViT-B/32	**34.75**	**11.00**	**7.50**
→ UOTScore (ours)	33.50	9.75	7.25
CLIPScore ViT-B/16	27.75	**10.75**	**8.25**
→ UOTScore (ours)	**28.00**	8.50	6.50
SigLIPScore ViT-B/16	**32.50**	12.50	**9.75**
→ UOTScore (ours)	31.00	**13.00**	8.50
SigLIP2Score ViT-B/16	**36.50**	**15.25**	**10.75**
→ UOTScore (ours)	31.75	10.00	7.75

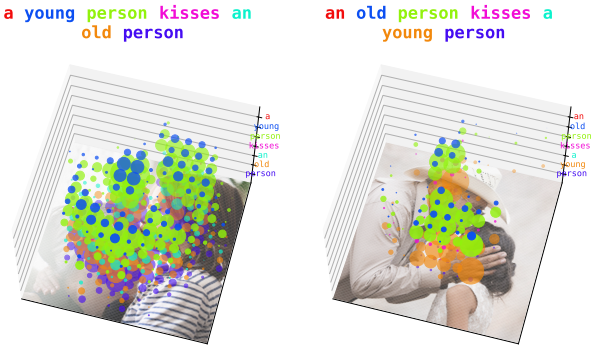

図 4: 正しいペアでの UOT の輸送計画

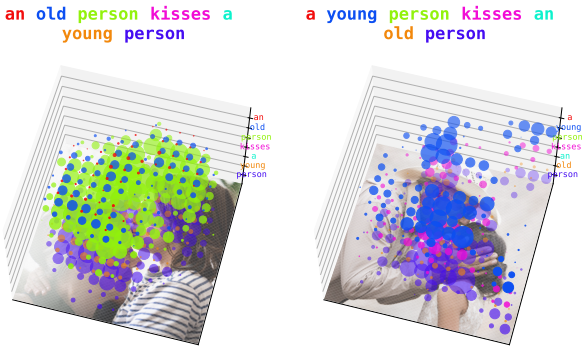

図 5: 誤ったペアでの UOT の輸送計画