複数粒度のマルチモーダル潜在表現に基づくネイルデザイン検索手法の構築と解析

木暮 緋南[1,a)]　雨宮 佳音[1]　杉浦 孔明[1,b)]

概要

本研究では，依頼文に基づきネイルデザイン画像を検索するタスクを扱う．依頼文には，自由な表現が可能なペイント部分および既製の装飾によるデコレーション部分から成るネイルデザインについて記述されており，多層的な要素を含むため，ユーザの意図に適合する画像の検索は困難である．既存手法は，依頼文の複雑な意図を十分に反映できず，一部の要素にのみ適合する画像を上位に順位付けする傾向がある．そこで本研究では，原画像および爪領域のみのマルチモーダル表現と，言語を媒介とした潜在表現との関係をモデル化することで，依頼文からネイルデザイン画像を検索するマルチモーダル検索手法を提案する．依頼文およびネイルデザイン画像から成る NAIL-STAR ベンチマークにおいて実験を行った．実験の結果，標準的な評価指標において，提案手法はベースライン手法を上回った．

1. はじめに

世界のネイルサロン市場規模は約 138 億米ドルであり [9]，ネイルデザインに対する社会的関心が高い．現状，ユーザは検索サイトを用いて希望するネイルデザインを検索することで，そのデザインを施術可能なネイリストを探すのが一般的である．従来のネイルデザインの検索には，掲載された画像に付与されたキーワードタグを用いる手法が用いられているが，これらのタグは必ずしも画像の内容を網羅的に記述しているとは限らず，ユーザが希望するデザインに合致する画像を検索することは困難である．そのため，ユーザが実際にネイリストに依頼するような自然言語の依頼文からネイルデザイン画像を検索できれば利便性が向上すると考えられる．なお，依頼文から画像を生成するアプローチも考えられるが，生成されたネイルデザインが再現可能とは限らないため，検索アプローチの方が現実的である．

本研究では，依頼文に基づきネイルデザイン画像を検索するタスク（図 1）を扱う．モデルは依頼文に記述された

図 1　本研究で扱うタスクの具体例

ユーザの意図に適合する画像を検索する．しかし，依頼文に記述された，詳細かつ多層的なユーザの意図に適合するネイルデザイン画像を検索することは困難である．これは，ネイルデザインが，自由な表現が可能であるペイント部分，および既製の装飾から選択および配置することで構成されるデコレーション部分から成るためである．

既存のマルチモーダル基盤モデル [13, 1] は，マルチモーダル画像検索タスク [7, 12] において良好な結果を得ているが，本タスクに直接適用するだけでは不十分である．具体的には，これらの手法は，依頼文に含まれる多層的な意図記述を十分に反映できず，その一部の要素にのみ適合する画像を優先的に上位に順位付けする傾向がある．さらに，雨宮ら [18] は画像全体から特徴量を統合しているため，ネイルデザイン以外の視覚要素の影響を受け，細部を捉えきれないという課題がある．

そこで，本研究では原画像に加え，片手の爪領域から得られる局所的なマルチモーダル特徴量を統合することで，依頼文からネイルデザイン画像を検索するマルチモーダル検索手法を提案する．これにより，細部の情報を捉えきれず，ネイルデザイン以外の視覚要素の影響を受けるという課題を解決する．既存手法との主要な違いは，原画像に加えて，爪領域を切り出した画像を用いて，マルチモーダル表現と言語を媒介とした潜在表現との関係性をモデル化する点にある．これらを導入することで，ネイルデザインの細部に着目した視覚特徴の活用が可能となり，背景などの

[1] 慶應義塾大学
[a)] hina.chick1717@keio.jp
[b)] komei.sugiura@keio.jp

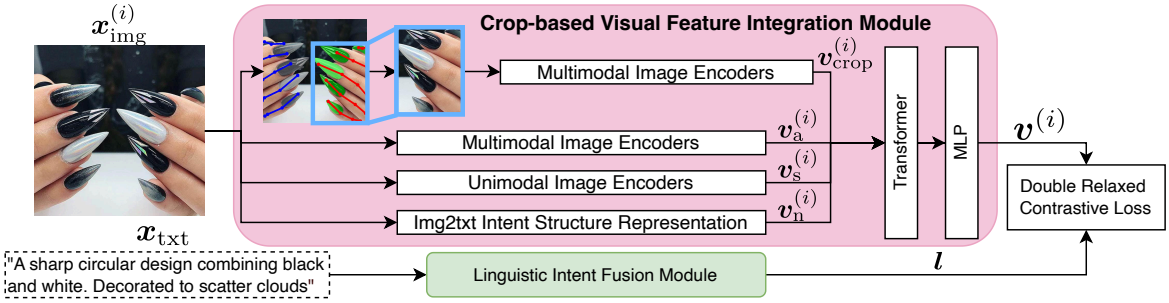

図 2 提案手法のモデル構造

不要な情報の影響が抑制されることが期待される．また，依頼文における複雑な参照関係を取り扱うことが可能となり，ユーザの意図により整合する画像の検索が可能になると考えられる．提案手法における新規性は次の通りである．

- 原画像および 爪領域のみのマルチモーダル表現と，言語を媒介とした潜在表現との間の関係をモデル化する Crop-based Visual Feature Integration Module (CVFIM) を導入する．これにより，ネイルデザインの細部の特徴を反映した複雑な参照関係を扱うことを可能とする．

2. 提案手法

提案手法のモデル構造を図 2 に示す．モデルへの入力を，$\boldsymbol{x} = \{\boldsymbol{x}_{\text{txt}}, X_{\text{img}}\}$，$X_{\text{img}} = \{\boldsymbol{x}_{\text{img}}^{(i)} \mid i = 1, \ldots, N_{\text{img}}\}$ と定義する．ここで，$\boldsymbol{x}_{\text{txt}} \in \{0,1\}^{V \times L}$ および $\boldsymbol{x}_{\text{img}}^{(i)} \in \mathbb{R}^{3 \times W \times H}$ はそれぞれ，トークナイズされた依頼文および，ネイルデザイン画像を表している．また，V，L，i，N_{img}，W および，H はそれぞれ，依頼文の語彙数，トークンの最大長，各画像のインデックス，順位付け対象となる画像の総数，画像の幅，画像の高さを表す．なお，$\boldsymbol{x}_{\text{txt}}$ に対しては，Lingistic Intent Fusion Module [18] を適用し，言語特徴量 $\boldsymbol{l}$ を抽出する．

CVFIM は，原画像および 爪領域のみのマルチモーダル表現と，言語を媒介とした潜在表現との間の関係をモデル化するモジュールである．既存手法 [18] では，両手が同一画像内に写っている場合でも $\boldsymbol{x}_{\text{img}}$ 全域に対して視覚特徴量を抽出していたため，各爪の占める面積が相対的に小さくなり，模様や装飾などのネイルデザインの細部情報が失われるという問題があった．また，背景などネイルデザイン以外の視覚要素により，依頼文との意味的な整合性を欠いた画像が検索される問題もあった．これらの課題を解決するため，本モジュールでは，爪領域を検出し，hand pose estimation によって推定された手の関節座標に基づいて各爪を左右に分けた上で，片手に属すると判断された複数の爪領域を含む矩形領域を切り出すアプローチを採用する．

まずセグメンテーションモデル [11] を用いて，$\boldsymbol{x}_{\text{img}}$ に対し各爪に対応するセグメンテーションマスクの集合 $\mathcal{M} = \left\{\boldsymbol{m}_j \in \{0,1\}^{H \times W}\right\}_{j=1}^{N_{\text{nail}}}$ を得る．N_{nail} は爪領域の検出数を表す．次に，hand pose estimation モデル（RTM-Pose [5]）を適用し，画像内の手の関節の 2 次元座標集合 $\{\boldsymbol{p}_k \in \mathbb{R}^2\}_{k=1}^{N_{\text{joint}}}$ を推定する．N_{joint} は関節の検出数を表す．続いて，$\boldsymbol{m}_j$ の重心を計算し，最も近い関節点に基づき，$\boldsymbol{m}_j$ を左手あるいは右手に分類する．分類された $\boldsymbol{m}_j$ の数が多い方の手を対象手 h^* として選択する．また，左右両方の手に分類された $\boldsymbol{m}_j$ の数が同数の場合，手に属するマスクの重心集合と画像中心との平均距離が短い手を h^* として選択する．次に，N_{nail}，および h^* に分類された $\boldsymbol{m}_j$ の数に応じて，切り出し画像 $\boldsymbol{x}_{\text{crop}}$ を生成する．

視覚特徴量は以下のようにして得られる．(1) $\boldsymbol{x}_{\text{crop}}$ および $\boldsymbol{x}_{\text{img}}$ に対してマルチモーダル画像エンコーダ（BEiT-3 [1]）を適用して，言語と整合された視覚特徴量 $\boldsymbol{v}_{\text{crop}}^{(i)} \in \mathbb{R}^{d_{\text{a}}}$ および $\boldsymbol{v}_{\text{a}}^{(i)} \in \mathbb{R}^{d_{\text{a}}}$ を得る．(2) $\boldsymbol{x}_{\text{img}}$ に対してユニモーダル画像エンコーダ（DINOv2 [10]）を適用して，ネイルデザインの色，形状，および質感に関する視覚特徴量 $\boldsymbol{v}_{\text{s}}^{(i)} \in \mathbb{R}^{d_{\text{s}}}$ を得る．(3) MLLM（GPT-4o[2] および Qwen2-VL[14]）によって生成された依頼文に対して，テキストエンコーダ（Stella [17]）を適用して，ネイルデザインの表象および複雑な参照関係を捉えた特徴量 $\boldsymbol{v}_{\text{n}}^{(i)} \in \mathbb{R}^{d_{\text{n}}}$ を得る．CVFIM における最終的な出力 $\boldsymbol{v}^{(i)} \in \mathbb{R}^{d_{\text{img}}}$ は以下の式で表される．

$$\boldsymbol{v}^{(i)} = \text{MLP}\left(\text{Transformer}\left(\left[\boldsymbol{v}_{\text{s}}^{(i)}; \boldsymbol{v}_{\text{a}}^{(i)}; \boldsymbol{v}_{\text{crop}}^{(i)}; \boldsymbol{v}_{\text{n}}^{(i)}\right]\right)\right)$$

ここで，MLP，および Transformer はそれぞれ，多層パーセプトロン，Transformer エンコーダを表す．

損失関数には，Double Relaxed Contrastive Loss [16] を用いた．推論時におけるモデルの出力は $\hat{Y}$ であり，これは，$\text{sim}\left(\boldsymbol{l}, \boldsymbol{v}^{(i)}\right)$ に基づいて X_{img} を降順に並べ替えたネイルデザイン画像のリストである．ここで，$\text{sim}\left(\boldsymbol{l}, \boldsymbol{v}^{(i)}\right)$ はコサイン類似度を表す．

3. 実験設定

本研究では，NAIL-STAR ベンチマーク [18] を使用した．本ベンチマークは，詳細かつ多層的にネイルデザインを説明した依頼文，および多様なネイルデザイン画像からなる一対一のペアによって構成されており，NAIL-STAR タスクにおいて標準的であることから採用した．ただし，本研究ではテスト集合の一部において，各依頼文に対して新たに複数の画像を正例としてラベル付けした．これは，元のテスト集合において各依頼文に対して 1 枚の画像のみに正例ラベルが付与されている一方，実際には複数のネイルデザイン画像が適合する場合があるためである．すなわ

表 1 ベースライン手法との定量的比較結果および Ablation Study の結果

[%]	複数目標画像				単一目標画像			
手法	MRR ↑	R@1 ↑	R@5 ↑	R@10 ↑	MRR ↑	R@1 ↑	R@5 ↑	R@10 ↑
CLIP [13] (fine-tuned)	25.09 (±0.24)	14.35 (±0.45)	35.70 (±0.81)	46.65 (±0.81)	24.23 (±0.41)	13.40 (±0.45)	34.13 (±0.60)	45.45 (±0.76)
BEiT-3 [1] (fine-tuned)	63.59 (±0.53)	50.83 (±0.68)	79.93 (±0.94)	87.70 (±0.75)	62.65 (±0.32)	50.18 (±0.52)	78.50 (±0.75)	86.18 (±0.82)
UniIR [15]	15.97	7.88	22.50	31.13	15.96	7.75	21.75	34.13
FashionViL [3] (frozen)	4.05	1.00	5.13	8.00	3.78	1.25	4.63	7.63
FashionViL [3] (fine-tuned)	28.47 (±0.57)	15.40 (±0.44)	42.45 (±0.86)	56.08 (±1.96)	29.69 (±0.92)	17.58 (±1.09)	41.13 (±0.54)	56.30 (±0.97)
FAME-ViL [4] (frozen)	26.59	15.50	38.88	52.25	26.47	16.25	36.25	48.88
FAME-ViL [4] (fine-tuned)	53.12 (±0.72)	38.90 (±0.93)	71.20 (±1.12)	81.48 (±0.84)	53.73 (±0.53)	40.08 (±0.82)	70.18 (±0.74)	81.33 (±0.30)
NaiLIA [18]	63.25 (±0.58)	50.73 (±0.71)	79.15 (±0.86)	87.05 (±1.21)	62.50 (±0.65)	50.60 (±0.94)	77.38 (±0.72)	85.65 (±0.69)
提案手法	64.36 (±0.46)	52.20 (±0.62)	79.68 (±1.23)	87.80 (±0.66)	63.20 (±0.93)	50.83 (±1.37)	78.85 (±0.61)	86.48 (±0.29)
提案手法 $+\boldsymbol{v}_{\text{crop}}$	**68.61** (±0.36)	**57.40** (±0.54)	**82.95** (±0.72)	**88.60** (±0.27)	**67.51** (±0.34)	**56.73** (±0.74)	**80.83** (±0.83)	**88.53** (±0.29)

ち，それらの画像も正例として扱われるべきであるが，従来のラベル付与方式では負例として扱われ，評価に反映されないという課題があった．

CVFIM で切り出し画像 $\boldsymbol{x}_{\text{crop}}$ を生成する際，N_{nail}，および h^* に分類された $\boldsymbol{m}_j$ の数に応じて処理を分岐した．以下に，$\boldsymbol{x}_{\text{crop}}$ の定義方法の分岐について示す．(1) $N_{\text{nail}} \leq 2$ の場合，適切な領域切り出しが困難と判断し，領域切り出し処理は行わず，$\boldsymbol{x}_{\text{crop}} = \boldsymbol{x}_{\text{img}}$ とした．(2) h^* に属するマスクの個数 N_{h^*} について，$N_{h^*} \leq 2$ の場合，左右両手の複数の爪領域を含むような矩形領域を $\boldsymbol{x}_{\text{crop}}$ とした．(3) $N_{h^*} = 3$ の場合，h^* に分類されなかった $\boldsymbol{m}_j$ の中から，h^* に属する爪領域の中心との距離が最も近いものを h^* に属するマスク集合に追加し，h^* に属する複数の爪領域を含む矩形領域を $\boldsymbol{x}_{\text{crop}}$ とした．(4) $N_{h^*} \geq 4$ の場合，h^* に属する複数の爪領域を含む矩形領域を $\boldsymbol{x}_{\text{crop}}$ とした．

4. 実験結果

4.1 定量的結果

本研究では，ベースライン手法として，UniIR [15]，FashionViL [3]，および FAME-ViL [4] を用いた．加えて，CLIP (ViT-B/32) [13]，BEiT-3 (large) [1]，FashionViL，FAME-ViL のテキストエンコーダと画像エンコーダの両方を fine-tuning したモデル，および NaiLIA [18] をベースライン手法とした．評価指標として，Mean Reciprocal Rank (MRR)，および Recall@K ($K = 1, 5, 10$) を用いた．主要評価指標は，Recall@10 とした [6, 8]．

表 1 に提案手法およびベースラインの定量的比較結果を示す．UniIR [15]，FahionViL [3]，および FAME-ViL [4] については，ゼロショット設定で実験を行った結果を示す．事前学習済みのモデルでは，複数回の試行において一貫した結果が得られるため，各手法につき 1 回の実験を実施した．fine-tuning したモデル，NaiLIA [18]，および提案手法については，5 回の実験を実施し，その結果の平均および標準偏差を示す．また，表中の太字は各評価指標における，最も高い数値を表す．

表 1 より，Recall@10 において，提案手法は複数目標画像サブセットで 88.60%，単一目標画像サブセットで 88.53% であった．一方，ベースライン手法の中で Recall@10 が最も高かったのは BEiT-3 を fine-tuning したモデルで，それぞれ 87.70%，および 86.18%であり，提案手法はこれらをそれぞれ 0.90 ポイント，および 2.35 ポイント上回った．また，両サブセットにおける他の評価指標においても，提案手法はベースライン手法を上回った．すべての評価指標において，ベースライン手法との性能差は統計有意であった（$p < 0.05$）．

図 3 提案手法およびベースライン手法 [1] の定性的結果

4.2 定性的結果

図 3 に，提案手法およびベースライン手法 [1] における定性的結果を示す．なお，今回は，ベースライン手法として Recall@10 が最も高かった BEiT-3 (fine-tuned) を採用し，複数目標画像サブセットにおける例を示す．ここでは，各手法における検索結果の上位 3 件の画像を示す．また，画像の緑枠は $\boldsymbol{x}_{\text{txt}}$ に対する目標画像群を表す．

図 3 の依頼文は “Please use long nail tips and marble pattern nails using bright colors.” である．提案手法は，目標画像群である，鮮やかな色彩を用いたマーブル模様の長いチップ形状のネイルデザイン画像を検索結果の 1 位と 3 位に順位付けした．一方，ベースライン手法では，上位 3 件においてマーブル模様のデザインが含まれず，さらに，色調が暗いものや短いチップ形状のデザインなど，依頼内

表 2　エラーのカテゴリおよびエラー数

Errors	#Error
色に関する表現の理解誤り	29
複合デザイン指示の理解誤り	15
テーマの理解誤り	12
イラストの理解誤り	11
曖昧性が大きい依頼文	9
ネイルパーツの理解誤り	6
模様の理解誤り	6
アノテーションエラー	5
爪の形や位置の理解誤り	3
その他	4
Total	100

容と一致しない画像が含まれていた．

4.3 Ablation Study

Ablation study として，CVFIM により得られた $\boldsymbol{x}_{\text{crop}}$ の性能への寄与を調べるため，$\boldsymbol{v}_{\text{crop}}$ を除いて実験を行なった．表 1 より，Recall@10 は，提案手法において複数目標画像サブセットで 87.80%，単一目標画像サブセットで 86.48%であり，提案手法 $+\boldsymbol{v}_{\text{crop}}$ と比較するとそれぞれ，0.80 ポイント，および 2.05 ポイント減少した．この結果より，CVFIM における $\boldsymbol{v}_{\text{crop}}$ は，ネイルデザイン以外の視覚情報の影響を抑えつつ，細部の特徴の反映を可能にすることで，性能に寄与していることが示唆される．

4.4 エラー分析

提案手法について，目標画像の順位が 11 位以下であった依頼文 100 文に対する検索結果のエラー分析を行った．表 2 に，エラーのカテゴリ，およびそれぞれのカテゴリにおけるエラー数を示す．各カテゴリは以下のように定義した．

(1) 色に関する表現の理解誤り: 色に関する依頼文の記述に対して，その意図を正しく理解できなかったエラーを指す．例えば，“a combination of blue-grey glitter nails and solid pink nails”という依頼文に対して，色の要素を分離的に処理し，青色と灰色とピンク色のそれぞれの色の指があるネイルデザインを上位に順位付けされた場合などを含む．

(2) 複合デザイン指示の理解誤り: 複数の視覚的要素の組み合わせや関係性が記述された依頼文に対し，個々の要素の役割や相互の配置，対応などを正しく理解できなかったエラーを指す．例えば，“Nails in a color close to gray with a slight glitter” という依頼文に対して，灰色とラメを個別の要素として理解し，灰色のネイルとラメネイルが別々に施された画像が上位に順位付けされた場合などを含む．

(3) テーマの理解誤り: ネイルデザイン全体に表現されたテーマという抽象的な意図を，正しく理解できなかったエラーを指す．ここで，テーマとは，特定の対象やイメージをもとに構想されたネイルデザインの方向性を指す．例えば，図 1 の依頼文およびネイルデザイン画像に示すように，花火や金魚といったイラストが描かれ，デザイン全体を通して日本の夏祭りを想起させるものなどが含まれる．

(4) イラストの理解誤り：ネイルデザインに含まれるイラストに関する表現を正しく理解できず，依頼文に含まれる意図と異なる図案を含むネイルデザイン画像を提示したエラーを指す．

(5) 曖昧性が大きい依頼文：依頼文の記述が抽象的または一般的であり，多数の画像が同程度に適合し得るケースを指す．

(6) ネイルパーツの理解誤り：固有のネイルパーツ名，およびネイルパーツの種類を正しく理解できなかったエラーを指す．ネイルパーツに関する依頼文の指定に対して，該当するパーツが存在しない，または異なる種類のパーツが用いられた画像が上位に順位付けされる場合などを含む．

(7) 模様の理解誤り：ネイルデザインにおける模様に関する依頼文の記述を正しく理解できず，意図とは異なる模様を含む画像を提示したエラーを指す．

(8) アノテーションエラー：依頼文の内容が目標画像と一致していなかったエラーを指す．

(9) 爪の形や位置の理解誤り：ネイルの長さや形，また，指の中での装飾の位置を正しく理解できなかったエラーを指す．例えば，長くて先端が尖っているネイルという指示に対して，短く，先端が尖っていないネイルデザイン画像が上位に順位付けされた場合などを含む．

(10) その他：上記のエラー以外を指す．

提案手法における主要なエラー要因は，色に関する表現の理解誤りであった．原因として，依頼文に記述された色を示す語，および画像において対応する色との整合性が十分でないことが挙げられる．解決策として，画像から色情報を直接抽出し，依頼文中に含まれる対象の色を表す語句と整合させるための損失関数を導入することが挙げられる．これにより，色に関する表現の理解に対する頑健性が向上することが期待される．

5. おわりに

依頼文に基づきネイルデザイン画像を検索する，NAIL-STAR タスクを扱った．本研究では，原画像および 爪領域のみのマルチモーダル表現と，言語を媒介とした潜在表現との間の関係をモデル化する CVFIM を導入した．また，NAIL-STAR ベンチマークで評価を行い，標準的な評価尺度において，提案手法がベースライン手法を上回った．将来研究として，色を明示的に示す単語に依存せず，依頼文全体の文脈から色に関する意図を推定する手法の検討が挙げられる．

謝辞

本研究の一部は，JSPS 科研費 23K28168 の助成を受けて実施されたものである．

参考文献

[1] *Image as a Foreign Language: BEIT Pretraining for Vision and Vision-Language Tasks* (2023).

[2] Achiam, J., Adler, S., Agarwal, S. et al.: GPT-4 Technical Report, *arXiv preprint arXiv:2303.08774* (2023).

[3] Han, X., Yu, L., Zhu, X., Zhang, L., Song, Y.-Z. and Xiang, T.: Fashionvil: Fashion-focused vision-and-language representation learning, *ECCV*, Springer, pp. 634–651 (2022).

[4] Han, X., Zhu, X., Yu, L., Zhang, L. et al.: FAME-ViL: Multi-Tasking Vision-Language Model for Heterogeneous Fashion Tasks, *CVPR*, pp. 2669–2680 (2023).

[5] Jiang, T., Lu, P., Zhang, L., Ma, N., Han, R., Lyu, C., Li, Y. and Chen, K.: Rtmpose: Real-time multi-person pose estimation based on mmpose, *arXiv preprint arXiv:2303.07399* (2023).

[6] Li, K., Zhang, Y., Li, K., Li, Y. and Fu, Y.: Visual semantic reasoning for image-text matching, *ICCV*, pp. 4654–4662 (2019).

[7] Lin, T., Maire, M., Belongie, S., Bourdev, L., Girshick, R. et al.: Microsoft COCO: Common Objects in Context, *ECCV*, pp. 740–755 (2014).

[8] Liu, T.-Y.: Learning to Rank for Information Retrieval, *Found. Trends Inf. Retr.*, Vol. 3, No. 3, pp. 225–331 (2009).

[9] Maximize Market Research: Nail Salon Market: Global Industry Analysis by Market Share, Trend, Size, Competitive Landscape, Regional Outlook and Forecast (2025–2032), https://www.maximizemarketresearch.com/market-report/nail-salon-market/195476/ (2025). Accessed: 2025-06-07.

[10] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A. et al.: Dinov2: Learning robust visual features without supervision, *arXiv preprint arXiv:2304.07193* (2023).

[11] PDumaid: nakhoon: Instance Segmentation Model, https://universe.roboflow.com/pdumaid-wuy6w/nakhoon (2025). Accessed: 2025-06-06.

[12] Plummer, B. et al.: Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models, *ICCV*, pp. 2641–2649 (2015).

[13] Radford, A., Kim, J. W., Hallacy, C. et al.: Learning Transferable Visual Models from Natural Language Supervision, *ICML*, pp. 8748–8763 (2021).

[14] Wang, P., Bai, S. et al.: Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution, *arXiv preprint arXiv:2409.12191* (2024).

[15] Wei, C., Chen, Y., Chen, H., Hu, H., Zhang, G., Fu, J., Ritter, A. and Chen, W.: Uniir: Training and benchmarking universal multimodal information retrievers, *ECCV*, Springer, pp. 387–404 (2024).

[16] Yashima, D. et al.: Open-Vocabulary Mobile Manipulation Based on Double Relaxed Contrastive Learning With Dense Labeling, *IEEE RA-L*, Vol. 10, No. 2, pp. 1728–1735 (2025).

[17] Zhang, D., Li, J., Zeng, Z. and Wang, F. W.: Jasper and Stella: distillation of SOTA embedding models , *arXiv preprint arXiv:2412.19048* (2024).

[18] 雨宮佳音，小松拓実，八島大地，是方諒介，　勝又圭，杉浦孔明：NaiLIA: 緩和損失に基づくネイルデザインのマルチモーダル検索，第 39 回人工知能学会全国大会 (2025).