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

NaiLIA: Multimodal Retrieval of Nail Designs Based on Relaxed Contrastive Loss

雨宮 佳音 Kanon Amemiya　小松 拓実 Takumi Komatsu　八島 大地 Daichi Yashima　是方 諒介 Ryosuke Korekata　勝又 圭 Kei Katsumata　杉浦 孔明 Komei Sugiura

慶應義塾大学
Keio University

We focus on the task of retrieving nail design images based on dense intent descriptions, which represent long and multi-layered user intent for nail designs. This is challenging because such descriptions specify flexibly created paintings and pre-manufactured embellishments, as well as visual characteristics, spatial relationships, higher-order themes, and overall impressions. Existing vision-and-language foundation models often struggle to capture the interplay between paintings and embellishments, failing to incorporate multi-layered intent descriptions. To address this, we propose NaiLIA, a method that enables the retrieval of nail design images that comprehensively align with descriptions with dense user intent. Our approach estimates confidence scores for images that align with a given description and can be considered as positive examples but are not explicitly labeled (unlabeled positives), and incorporates this score into the loss function. To evaluate NaiLIA, we constructed a benchmark consisting of 10,625 images collected from people with diverse cultural backgrounds. The images were annotated with long and dense intent descriptions given by over 200 annotators. Experimental results demonstrate that the proposed method outperforms standard methods by 20.9 points in terms of recall@10.

1. はじめに

ネイルサロンの世界市場規模は約 110 億ドルと評価され [Grand View Research]，ユーザの要望を満たすネイルデザイン，およびそのデザインを施術可能なネイリストの検索の需要は大きい．ネイルサロン利用者がネイリストにデザインを依頼する際，色や模様，質感などのペイント要素，およびネイルパーツなどのデコレーション要素に加え，各要素の組み合わせを通して表現されるモチーフや印象などを伝える．このようにユーザが多層的な意図を持つ場合，ネイリストへ依頼する際と同様の表現を用いてネイルデザイン画像を検索できれば便利である．しかし，ユーザの意図が詳細かつ多層的に表現された依頼文から，意図に適合するネイルデザインを検索することは困難である．その理由として，ネイルデザインは自由な表現が可能なペイント部分，および既製の装飾を選択，配置することのみ可能なデコレーション部分から構成されることが挙げられる．また，ユーザの依頼文には，視覚的な情報だけでなく，モチーフや印象に関する表現が含まれるためである．

本研究では，ネイルデザインの依頼文をもとに，依頼文に含まれる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タスクを扱う．図 1 に，本タスクの具体例を示す．いま，ユーザから “I want nails with a mermaid theme, using light blue as the base color. Please draw a mermaid fin on the middle finger and a seashell on the ring finger. Add a pearl nail accessory to the seashell. I’d like a fresh, glossy, and shiny look.” という依頼文が与えられたとする．このとき，図 1 の左下に大きく示す画像が上位の結果として検索されることが望ましい．当該画像のネイルデザインには，中指にヒレ，薬指に貝殻が描かれており，貝殻のデザインには複数のパールのネイルパーツが配置されていることから，人魚をモチーフとしていることが示唆される．また，水色のベースカラーは爽やかな印象を与え，ラメにより艶やかで光沢感のある仕上がりとなっている．モデルはこの画像を，依頼文に適合する他のネイルデザイン画像とともに，上位の検索結果としてユーザに提示する．

図 1: 本研究で扱うタスクの具体例

本タスクに関連が深いマルチモーダル検索分野では多くの研究が行われているが，それらの多くは，本タスクにおいて十分な性能を発揮できていない（4 節参照）．主な要因は，正例以外の全てのサンプルを負例として扱う InfoNCE 損失 [Oord 18] を用いた学習に依存している点にある．これらの手法は，特定の抽象度に対応するネイルデザイン画像にのみ高い類似度を与える傾向がある．例えば，‘flower nail parts’ は，写実的な花の装飾や，花のシルエットの金属製の装飾，花を模したキャラクターの装飾などを指すことがある．既存手法ではしばしば，特定の抽象度（例：写実的な装飾）のネイルデザインが上位に偏った検索結果となる．

そこで，本研究では，ユーザの意図が詳細かつ多層的に記述された依頼文に基づき，包括的に適合する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，NaiLIA を提案する．本研究の貢献は次の通りである．

- 正例としてラベル付けされていないが依頼文に適合する画

連絡先: 雨宮佳音，慶應義塾大学，神奈川県横浜市港北区日吉 3-14-1，kanon-amemiya@keio.jp

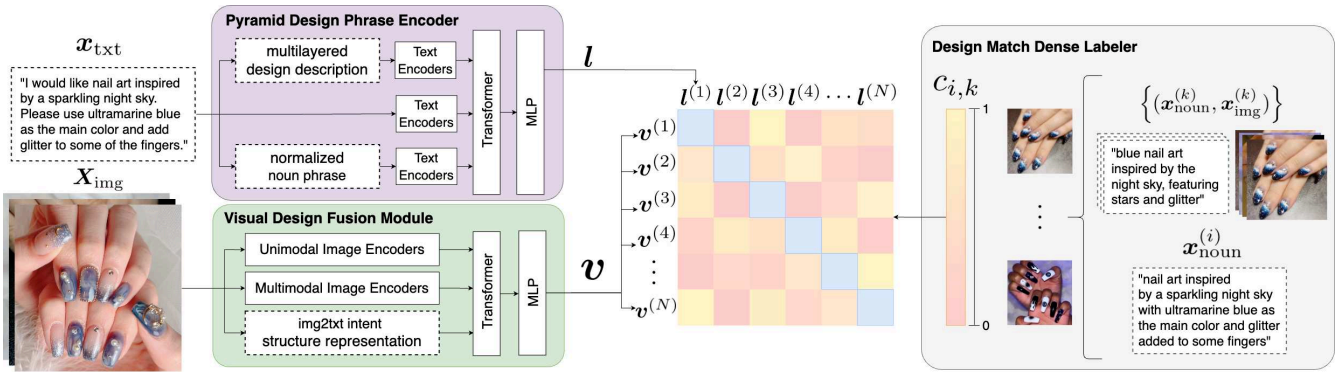

図 2: 提案手法のモデル構造

像（unlabeled positive）に対して確信度を推定し，これを損失に組み込むことで，unlabeled positive を考慮した学習を実現する Design Match Dense Labeler (DMDL) を導入する．

- ネイルデザイン画像から，(i) 色や形状などの視覚的な特徴量，(ii) 自然言語に整合された特徴量に加え，(iii) 言語を媒介することでデザインの表象や複雑な参照関係を捉えた特徴量を得て，これらを統合する Visual Design Fusion Module (VDFM) を導入する．
- 依頼文に含まれる多層的なユーザの意図を理解するため，元の依頼文を標準化することに加え，意図を構造化した文章を用いて，依頼文を階層構造としてモデル化する Pyramid Design Phrase Encoder (PDPE) を導入する．
- 詳細かつ多層的なユーザの意図の説明を伴う依頼文，および多様なネイルデザイン画像で構成された新規データセット Lunula Bench を構築する．

2. 問題設定

本研究では，ユーザによるネイルデザインの依頼文をもとに，ネイルデザイン画像を検索するタスクを扱う．本タスクでは，依頼文に表現されたユーザの意図に適合するネイルデザイン画像が，出力の画像リストにおいて高い順位として検索されることが望ましい．本タスクの入力は，ネイルデザインの依頼文およびネイルデザイン画像群であり，出力は，依頼文との関連性に基づきランク付けされたネイルデザイン画像群である．本論文では，正例ラベルがついており依頼文に適合するネイルデザイン画像を目標画像，依頼文に適合しているが，明示的にラベル付けされていないネイルデザイン画像を unlabeled positive と定義する．本研究では，ペイントやデコレーションなどが施された手の爪に焦点を当てた画像のみを扱う．

3. 提案手法

本研究では，マルチモーダル検索手法を拡張し，依頼文をもとにネイルデザイン画像を検索する NaiLIA を提案する．図 2 に，提案手法のモデル構造を示す．提案手法は，PDPE，VDFM，および DMDL の 3 モジュールから構成される．モデルへの入力 $\boldsymbol{x}$ は，以下の式で定義される．

$$\boldsymbol{x} = \{\boldsymbol{x}_{\text{txt}}, X_{\text{img}}\}, \quad X_{\text{img}} = \{\boldsymbol{x}^{(i)}_{\text{img}} \mid i = 1, \ldots, N_{\text{img}}\},$$

ここで，$\boldsymbol{x}_{\text{txt}} \in \{0,1\}^{V \times L}$ および $\boldsymbol{x}^{(i)}_{\text{img}} \in \mathbb{R}^{3 \times W \times H}$ 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す．また，V，L，i，N_{img}，W，および H はそれぞれ，依頼文の語彙サイズ，最大トークン長，各画像のインデックス，ランク付け対象の画像数，画像の幅，画像の高さを表す．

3.1 Pyramid Design Phrase Encoder

PDPE では，依頼文に含まれる多層的なユーザの意図を理解するため，元の依頼文を標準的な文型に変換することに加え，依頼文の意図を構造化した文章を生成し，依頼文を階層構造としてモデル化する．ネイルデザインの依頼文は通常，ペイントやデコレーション，爪の形状などの視覚的情報に加え，それらから連想されるモチーフ，およびデザインに対する印象などから構成される．依頼文はネイルサロン利用者がネイリストにデザインを依頼する際の表現を用いているため，抽象度の異なる要望が混在し，明瞭性に欠ける場合や冗長性が高い場合がある．このような依頼文からユーザの意図を適切に理解するため，以下の 2 つの文を生成し，これらを階層構造として扱う．PDPE の入力は $\boldsymbol{x}_{\text{txt}}$ である．$\boldsymbol{x}_{\text{txt}}$ は，(1) ペイントやデコレーション，(2) 爪の形状，(3) モチーフ，および (4) 印象の 4 つの項目に分割できる．よって，大規模言語モデル（GPT-4o）を用いて $\boldsymbol{x}_{\text{txt}}$ を 当該の 4 つの項目に分けて構造化した $\boldsymbol{x}_{\text{mdd}}$ を生成する．次に，冗長性を排除し，要点を明確にした文を得るために，$\boldsymbol{x}_{\text{txt}}$ が示すネイルデザインを名詞句に換言した $\boldsymbol{x}_{\text{nnp}}$ を得る．続いて，$\boldsymbol{x}_{\text{txt}}$，$\boldsymbol{x}_{\text{mdd}}$，および $\boldsymbol{x}_{\text{nnp}}$ に対してそれぞれ，複数のテキストエンコーダ（BEiT-3 [Wang 23], Stella [Zhang 24b]）を用いて，3 層の構造化された特徴量 $\boldsymbol{l}_{\text{txt}}, \boldsymbol{l}_{\text{mdd}}, \boldsymbol{l}_{\text{nnp}} \in \mathbb{R}^{d_{\text{txt}}}$ を得る．ここで，d_{tex} はテキストエンコーダの出力次元を表す．PDPE における最終的な出力 $\boldsymbol{l} \in \mathbb{R}^{d_{\text{txt}}}$ は以下の式で得られる．

$$\boldsymbol{l} = \text{MLP}\left(\text{Transformer}\left([\boldsymbol{l}_{\text{txt}}; \boldsymbol{l}_{\text{mdd}}; \boldsymbol{l}_{\text{nnp}}]\right)\right)$$

ここで，MLP(·)，Transformer(·)，および d_{txt} はそれぞれ，多層パーセプトロン，Transformer エンコーダ，および PDPE の出力次元を表す．

3.2 Visual Design Fusion Module

VDFM では，λ-Representation Encoder [Goko 24] を拡張し，ネイルデザイン画像から色や形状などの視覚的な特徴量，自然言語に整合された特徴量，デザインの表象や複雑な参照関係を捉えた特徴量を得て，これらを統合する．図 1 の左側に表示された画像に示すネイルデザインは人魚をモチーフとしており，中指にヒレ，薬指に貝殻が描かれているが，画像エンコーダを直接適用するだけでは，実体ではないモチーフや，指とデザインの対応関係に関する情報を含む特徴量抽出が不十分な場合がある．そこで，VDFM では，ユニモーダル画像エンコーダおよびマルチモーダル基盤モデルの画像エンコーダから得た特徴量に加え，マルチモーダル大規模言語モデル（MLLM）由来の言語を媒介とした特徴量を統合することで，ネイルデザイン画像の包括的な特徴量を取得する．VDFM の入力は $\boldsymbol{x}^{(i)}_{\text{img}}$ である．まず，ネイルデザインの色や形状，質感に関する特徴量を得るため，ユニモーダル画像エンコーダ（DINOv2 [Darcet

24]）を用いて $\boldsymbol{x}_{\text{img}}^{(i)}$ から特徴量 $\boldsymbol{v}_{\text{s}}^{(i)} \in \mathbb{R}^{d_{\text{s}}}$ を抽出する．また，自然言語と整合されたマルチモーダル特徴量を得るため，マルチモーダル画像エンコーダ（BEiT-3）を用いて $\boldsymbol{x}_{\text{img}}^{(i)}$ から特徴量 $\boldsymbol{v}_{\text{a}}^{(i)} \in \mathbb{R}^{d_{\text{a}}}$ を抽出する．ここで，d_{s} および d_{a} は，画像エンコーダの出力次元を表す．次に，複数の MLLM（GPT-4o, Qwen2-VL [Wang 24]）を用いて，$\boldsymbol{x}_{\text{img}}^{(i)}$ についてネイルデザインに注目した説明文を生成する．続いて，テキストエンコーダ（Stella）を用いて，生成した説明文から，デザインの表象や複雑な参照関係を捉えた特徴量を得る．ここで，d_{n} は，テキストエンコーダの出力次元を表す．VDFM における最終的な出力 $\boldsymbol{v}^{(i)} \in \mathbb{R}^{d_{\text{img}}}$ は以下の式で得られる．

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

ここで，d_{img} は VDFM の出力次元を表す．

3.3 Design Match Dense Labeler

DMDL は，unlabeled positive を考慮した学習を実現するため，unlabeled positive に対して確信度を推定し，これを利用して損失の計算を行う．既存のマルチモーダル対照学習手法では，主に InfoNCE [Oord 18] のような対照損失が使用される [Radford 21, Sun 24, Zhang 24a]．InfoNCE を用いた学習では，単一のテキストに対して単一の画像とのペアのみを正例とし，他の画像とのペアをすべて負例として扱い，正例間の類似度を最大化，負例間の類似度を最小化するようにモデルを最適化する．この方法では，バッチ内に正例とみなせるサンプルが存在する場合にも，そのサンプルを負例として扱う．そのため，類似度を最大化すべきであるとみなせるペアについて，類似度を最小化する形で学習を行うことから，一対一のラベル付けによる学習はノイズが生じやすい．そこで，本研究では unlabeled positive の候補画像に対して，unlabeled positive とみなせる程度を表す確信度を推定し，これをもとに unlabeled positive を考慮する損失関数を導入する．

訓練集合中のすべてのペアに対して確信度の推定を行う場合，訓練集合中の (文数 × 画像数) 回の推定を行う必要があるため，unlabeled positive である可能性の高いサンプルの集合に絞って推定を行うことが望ましい．よって，はじめに既存の視覚言語基盤モデル（BEiT-3）を用いて，$\boldsymbol{x}_{\text{txt}}^{(i)}$ および $\boldsymbol{x}_{\text{img}}^{(j)}$ から言語特徴量 $\boldsymbol{l}'^{(i)}$ および視覚特徴量 $\boldsymbol{v}'^{(j)}$ を抽出し，類似度 $\text{sim}\left(\boldsymbol{l}'^{(i)}, \boldsymbol{v}'^{(j)}\right) \in [-1, 1] (i \neq j)$ を計算する．ここで，$\boldsymbol{x}_{\text{txt}}^{(i)}$ との類似度の高い $\boldsymbol{x}_{\text{img}}^{(j)}$ が unlabeled positive である可能性が高いことから，類似度の上位 N_{cand} 枚の画像 $\left\{\boldsymbol{x}_{\text{img}}^{(k)}\right\}$（$k$ は上位 N_{cand} 枚の画像のインデックス集合の各要素）を unlabeled positive の候補画像群とする．次に，以下の式で示すように，$\boldsymbol{x}_{\text{img}}^{(k)}$ が $\boldsymbol{x}_{\text{txt}}^{(i)}$ の unlabeled positive とみなせる程度を表す確信度 $c_{i,k} \in [0, 1]$ を，MLLM（Qwen2-VL）を用いて推定する．

$$c_{i,k} = f(\boldsymbol{x}_{\text{nnp}}^{(i)}, \boldsymbol{x}_{\text{img}}^{(k)}, \boldsymbol{x}_{\text{nnp}}^{(k)}, \boldsymbol{x}_{\text{prompt}})$$

ここで，$\boldsymbol{x}_{\text{nnp}}^{(i)}$ および $\boldsymbol{x}_{\text{img}}^{(k)}$ だけでなく，$\boldsymbol{x}_{\text{nnp}}^{(k)}$ を用いる理由は，デザインの差分に着目させるためである．実際，$\boldsymbol{x}_{\text{nnp}}^{(i)}$ および $\boldsymbol{x}_{\text{img}}^{(k)}$ のみを入力とした場合，いずれも爪に焦点を当てた記述および画像であることに起因して，大きく異なるネイルデザインの画像に対しても不当に高い値を出力するという問題がある．$\boldsymbol{x}_{\text{nnp}}^{(k)}$ を参照文として用いることにより，デザイン同士の差分を言語情報として明確にすることで，大きく異なるデザインに対して高いスコアを与えることを抑制する．この方法は，同一カテゴリの物体を扱う問題設定に限らず，類似画像を多数含む他のマルチモーダル検索タスクにも広く適用可能である．

$c_{i,k} \geq \theta$（θ は閾値）の場合，$\boldsymbol{x}_{\text{img}}^{(k)}$ は $\boldsymbol{x}_{\text{txt}}^{(i)}$ に対する unlabeled positive として，unlabeled positive の集合 $\mathcal{S}$ に (i, k) を追加する．実験では，MLLM の出力は ‘0’, ‘5’, または ‘10’ とし，それぞれスケールした値 0，0.5，および 1 を c とする．損失関数には Double Relaxed Contrastive loss [Yashima 25] を用いる．

4. 実験

4.1 Lunula Bench

本研究では，ネイルデザインを詳細かつ多層的に説明した依頼文，および多様なネイルデザイン画像の 10,625 組のペアから構成される Lunula Bench を新規に構築した．はじめに，Pinterest [*1] からネイルデザイン画像を収集した．続いて，(1) 足のネイルデザインやネイルチップの画像 (2) 画像全体に対する爪の面積が極端に小さい画像 (3) 爪の数が 2 以下または 11 以上の画像 (4) 重複画像 (5) 生成 AI による画像 (6) 照明が極端に暗くデザインの認識が困難な画像を除外した．次に，上記のフィルタリングにおいて残った画像に対して，ネイルデザインを詳細かつ多層的に説明した依頼文のアノテーションを行った．アノテータには，表示されたネイルデザイン画像に対して，“If you were to request this nail design from a nail artist, how would you request?” という質問に，10 単語以上の英語の文章で回答するよう指示した．この際，各ネイルデザインについて，色，模様，質感，ネイルパーツなどの要素を含む視覚的な情報を記述するよう指示した．また，視覚的な情報が示唆するモチーフや，デザイン全体が与える印象も併せて記述するよう求めた．単に画像キャプショニングを行うのではなく，ネイリストに対して依頼を行う場面を想定した上で回答するよう注意した．208 人のアノテータから，合計 10,625 件の依頼文を収集した．同じ依頼文を連続して入力したアノテータや，応答時間が極端に短いアノテータのデータは，データセットの品質向上のため除外した．本データセットは，訓練集合，検証集合，テスト集合としてそれぞれ，8,625 サンプル，400 サンプル，1,600 サンプルに分割した．訓練集合はモデルのパラメータ更新，検証集合はハイパーパラメータの調整に使用した．また，テスト集合はモデルの評価に使用した．

Lunula Bench の新規性は次の通りである．(1) 本データセットのネイルデザイン画像は，自由な表現が可能なペイント部分，および既製品のネイルパーツを選択，配置することのみ可能なデコレーション部分から構成される．したがって，単色のみで塗られた爪画像から構成される既存の爪画像を扱うデータセットとは異なる．また，プロンプトと生成画像のペアや，既製品のみで構成されたデータセットとも異なる．(2) 本データセットの依頼文は，視覚的情報に加えて，デザインから連想されるモチーフやデザインに対する印象など，ユーザの意図が多層的に記述されている．また，特定の指に対して色を指定したり，特定の色に対して模様を指定したりなど，各要素が複雑な対応関係を持つ．

4.2 実験設定

本研究では，ベースライン手法として，CLIP (ViT-B/32) [Radford 21], BLIP-2 (ViT-g) [Li 23], BEiT-3 (large) [Wang 23], Alpha-CLIP (ViT-L/14) [Sun 24], および Long-CLIP (ViT-L/14) [Zhang 24a] を用いた．さらに，CLIP（ViT-B/32）のテキストエンコーダおよび画像エンコーダを fine-tuning したモデルを用いた．CLIP, BLIP-2, BEiT-3, および Long-CLIP は，ゼロショット設定の text-to-image

*1 https://pinterest.com/

検索タスクにおける代表的な手法であるため選択した．Alpha-CLIP はマスクを活用することで，ネイルに焦点を当てたマルチモーダル検索を可能とするため選択した [Sun 24]．評価尺度には，mean reciprocal rank（MRR）および recall@10 を用いた．訓練には 24GB の GPU メモリ搭載の GeForce RTX 4090，Intel Core i9-13900KF，および 64GB の RAM を用いた．モデルの訓練時間は約 20 分，1 つの依頼文に対して 100 枚の画像の類似度計算に要する推論時間は約 0.37 秒であった．各エポック毎に検証集合を用いて recall@10 を算出し，最も高いスコアを得たモデルを用いてテスト集合で評価した．

表 1: ベースライン手法との定量的結果

[%]	手法	MRR ↑	R@10 ↑
(i)	CLIP (freezed) [Radford 21]	12.4	23.7
(ii)	CLIP (fine-tuned) [Radford 21]	18.2	35.8
(iii)	BLIP-2 [Li 23]	14.4	28.0
(iv)	BEiT-3 [Wang 23]	34.9	57.9
(v)	Alpha-CLIP [Sun 24]	19.6	34.3
(vi)	Long-CLIP [Zhang 24a]	10.6	19.7
(vii)	NaiLIA (ours)	**54.7**	**78.8**

4.3 定量的結果

表 1 に，ベースライン手法と提案手法の定量的結果を示す．Alpha-CLIP については，依頼文およびネイルデザイン画像に加え，ネイルデザイン画像から生成した爪のセグメンテーションマスクを入力した．表中の太字は各評価尺度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 (vii) は 78.8%であり，ベースライン (i)，(ii)，(iii)，(iv)，(v)，(vi) をそれぞれ 55.1，43.0，50.8，20.9，44.5，59.1 ポイント上回った．また，提案手法は他の評価尺度においても，ベースライン手法を上回った．

4.4 定性的結果

図 3 に，提案手法およびベースライン手法 [Wang 23] における定性的結果を示す．ここでは目標画像，および各手法における上位 3 件の画像を示す．また，画像の青色，黄色，赤色の枠はそれぞれ，$\boldsymbol{x}_{\text{txt}}$ に対する正例，unlabeled positive，負例を表す．依頼文は “I’d like a colorful and flashy nail design. Please add a large flower nail stone to the ring finger. The tips of the nail tips should be square-shaped.” である．ここで，依頼文に含まれる ‘a large flower nail stone’ について，ユーザは実際の花と酷似した外観の装飾を意図しているとは限らず，目標画像に示すように，花を模したキャラクターの装飾などを意図している場合がある．目標画像における花を模した装飾は，実際の花とは外観が大きく乖離しているが，提案手法はこの装飾が花を象徴していることを正しく認識し，目標画像を 1 位として検索した．また，薬指の指定には適合していない反面，依頼文に含まれるネイルパーツや爪の形状の指定に加えて，‘coloful and flashy’ というデザインに対する印象の条件を満たすネイルデザイン画像が，unlabeled positive として 3 位として検索された．一方，ベースライン手法は目標画像を 137 位とし，花のネイルパーツが装飾されているネイルデザインを上位 3 位に 1 つも含まなかった．

5. おわりに

本研究では，ネイルデザインの依頼文をもとに，依頼文の要求に適合するネイルデザイン画像を検索するタスクを扱った．実験の結果，マルチモーダル検索タスクの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 20.9 ポイント上回った．将来研究としては，PDPE の各層における表現の違いを活用するため，各層に対応した損失関数を設計することが挙げられる．

$\boldsymbol{x}_{\text{txt}}$: “I'd like a colorful and flashy nail design. Please add a large flower nail stone to the ring finger. The tips of the nail tips should be square-shaped.”

図 3: 提案手法およびベースライン手法 [Wang 23] の定性的結果

謝辞

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

参考文献

[Darcet 24] Darcet, T., Oquab, M., Mairal, J., et al.: Vision Transformers Need Registers, in *ICLR* (2024)

[Goko 24] Goko, M., Kambara, M., et al.: Task Success Prediction for Open-Vocabulary Manipulation Based on Multi-Level Aligned Representations, in *CoRL* (2024)

[Grand View Research] Grand View Research, : Nail Salon Market Size & Trends, `https://www.grandviewresearch.com/industry-analysis/nail-salon-market-report`: Accessed: 2025-02-12

[Li 23] Li, J., et al.: BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, in *ICML*, Vol. 202, pp. 19730–19742 (2023)

[Oord 18] Oord, A., Li, Y., and Vinyals, O.: Representation Learning with Contrastive Predictive Coding, *arXiv preprint arXiv:1807.03748* (2018)

[Radford 21] Radford, A., Kim, J. W., et al.: Learning Transferable Visual Models From Natural Language Supervision, in *ICML*, pp. 8748–8763 (2021)

[Sun 24] Sun, Z., Fang, Y., Wu, T., Zhang, P., et al.: Alpha-CLIP: A CLIP Model Focusing on Wherever You Want, in *CVPR*, pp. 13019–13029 (2024)

[Wang 23] Wang, W., et al.: Image as a Foreign Language: BEiT Pretraining for Vision and Vision-Language Tasks, in *CVPR*, pp. 19175–19186 (2023)

[Wang 24] 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)

[Yashima 25] 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)

[Zhang 24a] Zhang, B., Zhang, P., Dong, X., Zang, Y., et al.: Long-CLIP: Unlocking the Long-Text Capability of CLIP, in *ECCV*, pp. 311–329 (2024)

[Zhang 24b] Zhang, D., Li, J., Zeng, Z., and Wang, F. W.: Jasper and Stella: distillation of SOTA embedding models , *arXiv preprint arXiv:2412.19048* (2024)