# NaiLIA: 多層的な依頼文に基づく ネイルデザインのマルチモーダル検索

雨宮 佳音[1,a)]　小松 拓実[1]　八島 大地[1]　是方 諒介[1]　勝又 圭[1]　杉浦 孔明[1,b)]

## 概要

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．依頼文には，ネイルデザインのペイント要素やデコレーション要素などの視覚的特徴に加え，空間的関係，テーマ，印象などユーザの多層的な意図が記述されている．既存手法では，このような複雑な依頼文から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．そこで，本研究では， 依頼文に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．また，NaiLIA を評価するため，1 万枚以上の多様なネイルデザイン画像，および各画像に対してアノテーションされた依頼文から構成されるデータセットを構築した．実験の結果，標準的な画像検索指標において，NaiLIA はベースライン手法を上回った．

## 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 [2]，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．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．ユーザから “I want nails with a mermaid theme, ... I’d like a fresh, glossy, and shiny look,” という依頼文が与えられたとき，図 1 の左下に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．

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

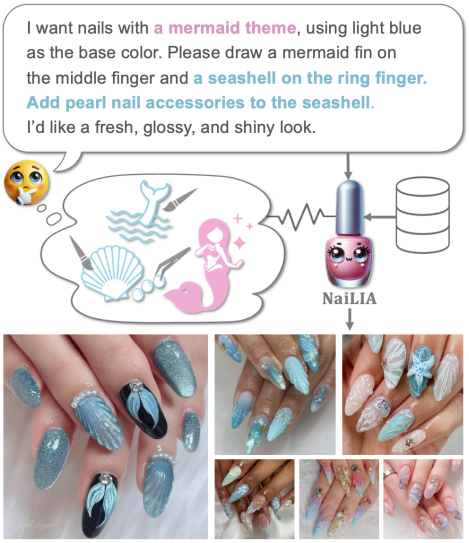


**図 1**　本研究で扱うタスクの具体例

は困難である．その理由は，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ためである．また，依頼文には色や模様などの視覚的情報に加え，モチーフや印象に関する表現が含まれるためである．

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 [4] 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像（unlabeled positive）に対して確信度を推定

[1] 慶應義塾大学
[a)] kanon-amemiya@keio.jp
[b)] komei.sugiura@keio.jp

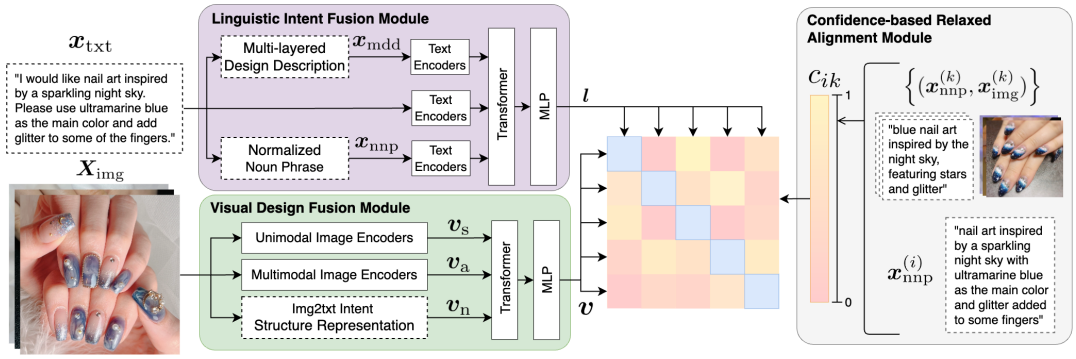


**図 2** 提案手法のモデル構造

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 Confidence-based Relaxed Alignment Module (CRAM) 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，依頼文から名詞句および意図を構造化した文章を得て，依頼文を階層構造としてモデル化する Linguistic Intent Fusion Module (LIFM) 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ベンチマーク NAIL-STAR を構築する．

## 2. 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，LIFM，VDFM，および CRAM の 3 モジュールから構成される．モデルへの入力 $\boldsymbol{x}$ は，$\boldsymbol{x} = \{\boldsymbol{x}_{\text{txt}}, X_{\text{img}}\}$，$X_{\text{img}} = \{\boldsymbol{x}_{\text{img}}^{(i)} \mid i = 1, \ldots, N_{\text{img}}\}$ と定義される．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す．また，$V$，$L$，$i$，$N_{\text{img}}$，$W$，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，検索対象の画像数，画像の幅，高さを表す．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．

### 2.1 Linguistic Intent Fusion Module

LIFM 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，モチーフや印象などに関する表現から構成される．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．LIFM 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 [8], Stella [10]）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，$d_{\text{tex}}$ はテキストエンコーダの出力次元を表す．LIFM 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$(\cdot)$，Transformer$(\cdot)$，および $d_{\text{txt}}$ はそれぞれ，多層パーセプトロン，Transformer エンコーダ，および LIFM の出力次元を表す．

### 2.2 Visual Design Fusion Module

VDFM では，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 $\boldsymbol{x}_{\text{img}}^{(i)}$ 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 [1]）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，$d_{\text{s}}$ および $d_{\text{a}}$ 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 [7]）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 $\boldsymbol{v}_{\text{n}}^{(i)} \in \mathbb{R}^{d_{\text{n}}}$ を得る．ここで，$d_{\text{n}}$ 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，$d_{\text{img}}$ は VDFM の出力次元を表す．

### 2.3 Confidence-based Relaxed Alignment Module

CRAM 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 [4] のような対照損失が使用される [5, 6, 9]．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う．そのため，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．

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 $N_{\text{cand}}$ 枚の画像 $\left\{\boldsymbol{x}_{\text{img}}^{(k)}\right\}$（$k$ は上位 $N_{\text{cand}}$ 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 $c_{ik} \in [0, 1]$ を，MLLM（Qwen2-VL）を用いて推定する．

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

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．$c_{ik} \geq \theta$（$\theta$ 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．

## 3. 実験

### 3.1 NAIL-STAR ベンチマーク

本研究では，ネイルデザインを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 NAIL-STAR ベンチマークを新規に構築した．ネイルデザイン画像は Pinterest [*1] から収集し，208 人のアノテータより依頼文を収集した．本ベンチマークは，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 サンプルから構成される．

NAIL-STAR ベンチマークの新規性は次の通りである．(1) 本ベンチマーク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) 本ベンチマークの依頼文は，視覚的情報に加えて，モチーフや印象に関する表現など，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．

### 3.2 実験設定

本研究では，ベースライン手法として，CLIP (ViT-B/32) [5], BLIP-2 (ViT-g) [3], BEiT-3 (large) [8], Alpha-CLIP (ViT-L/14) [6]，および Long-CLIP (ViT-L/14) [9] 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 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 [6]．評価尺度には，mean reciprocal rank（MRR）および recall@10 を用いた．

### 3.3 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 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

[*1] https://pinterest.com/

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) [5] | 12.4 | 23.7 |
| (ii) | CLIP (fine-tuned) [5] | 18.2 | 35.8 |
| (iii) | BLIP-2 [3] | 14.4 | 28.0 |
| (iv) | BEiT-3 [8] | 34.9 | 57.9 |
| (v) | Alpha-CLIP [6] | 19.6 | 34.3 |
| (vi) | Long-CLIP [9] | 10.6 | 19.7 |
| (vii) | NaiLIA (ours) | **54.7** | **78.8** |

(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．

### 3.4　定性的結果

図 3 に，提案手法およびベースライン手法 [8] 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．画像の緑および黄色の枠はそれぞれ，$\boldsymbol{x}_{\text{txt}}$ に対する正例および unlabeled positive を表す．(a) に提案手法における成功例を示す．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，“colorful 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．

(b) には提案手法において，目標画像が 1 位，unlabeled positive が 2 位として検索された成功例を示す．提案手法は，ステンドガラス調のデザインが，目標画像に示すような一部にラメが塗られた幾何学模様のデザインであると判断し，類似したデザインを上位 2 位にランク付けした．一方，ベースライン手法では依頼文に含まれるユーザの意図にほとんど適合しないネイルデザイン画像を上位 3 位に，目標画像を 5 位にランク付けした．

### 3.5　Ablation Study

**表 2** に，ablation study の結果を示す．ablation 条件として以下を定めた．

**LIFM ablation.** PDPE において，Transformer エンコーダおよび多層パーセプトロンの構造を維持したまま，LIFM から $\boldsymbol{l}_{\text{mdd}}$ および $\boldsymbol{l}_{\text{nnp}}$ を除去し，それぞれの寄与を調査した．Recall@10 において，モデル (b) および (c) はモデル (a) と比較してそれぞれ，1.5 ポイントおよび 2.1 ポイント低下した．この結果は，依頼文を構造化した文章および名詞句の，双方を組み合わせて用いることの有効性を示す．

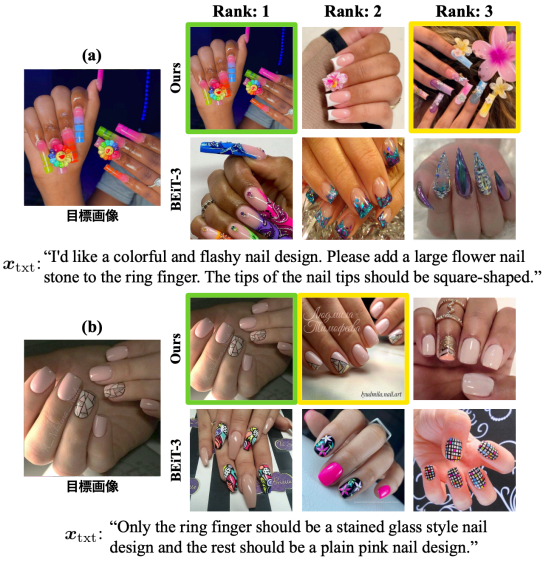


図 3　提案手法およびベースライン手法 [8] の定性的結果

表 2　ablation study における定量的結果

| [%] | モデル | MRR ↑ | R@10 ↑ |
|---|---|---|---|
| (a) | NaiLIA (full) | **55.1** | **78.8** |
| (b) | w/o $\boldsymbol{l}_{\text{mdd}}$ | 52.9 | 77.3 |
| (c) | w/o $\boldsymbol{l}_{\text{nnp}}$ | 52.8 | 76.7 |
| (d) | w/o $\boldsymbol{v}_{\text{s}}$ | 54.7 | 78.6 |
| (e) | w/o $\boldsymbol{v}_{\text{a}}$ | 48.8 | 73.7 |
| (f) | w/o $\boldsymbol{v}_{\text{n}}$ | 50.1 | 75.0 |
| (g) | w/ InfoNCE [4] | 52.6 | 77.4 |
| (h) | $c_{i,j} = 0.7$ | 54.8 | 78.5 |

**VDFM ablation.** 同様に，VDFM から $\boldsymbol{v}_{\text{s}}$, $\boldsymbol{v}_{\text{a}}$, および $\boldsymbol{v}_{\text{n}}$ をそれぞれ除去し，各特徴量の寄与を調査した．Recall@10 において，モデル (d)，(e)，および (f) はモデル (a) と比較してそれぞれ，0.2 ポイント，5.1 ポイント，および 3.8 ポイント低下した．この結果は，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，および言語を媒介として構造化された特徴量の 3 種類の潜在表現が，それぞれ補完的な関係にあることを示唆する．

**CRAM ablation.** 損失関数を InfoNCE[4] に変更し，unlabeled positive を考慮した学習の有効性を調査した．さらに，MLLM によるスコア推定の妥当性を評価するため，推定値の代わりに $c_{ik} = 0.7$ に固定する実験を行った．Recall@10 において，モデル (g) および (h) モデル (a) と比較してそれぞれ，1.4 ポイントおよび 0.3 ポイント低下した．この結果は，unlabeled positive を考慮できる損失関数の導入と，信頼度スコアの推定が，効率的な学習に寄与することを示唆する．

## 4.　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．

**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


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

[2] Grand View Research: Nail Salon Market Size & Trends, https://www.grandviewresearch.com/industry-analysis/nail-salon-market-report (2023).

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

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

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

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

[7] 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).

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

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

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