NaiLIA: Multimodal Nail Design Retrieval Based on Dense Intent Descriptions and Palette Queries

Supplementary Material

In this supplementary material, we provide the prompts and generated texts used in each module of NaiLIA, along with details of the NAIL-STAR benchmark construction and data distributions. We also present additional quantitative and qualitative results, including ablation studies and error analyses. Furthermore, the supplementary material includes a project page, which readers may refer to for an overview of NaiLIA or to view qualitative results at a higher resolution.

A. Additional Related Work

Fashion AI. In the domain of fashion AI, numerous studies have been conducted on various tasks, including clothing category recognition [59], fashion image retrieval [27, 46], virtual try-on [4, 13, 43], fashion recommendation [68], trend forecasting [1], and fashion compatibility [45]. Representative subfields in fashion AI include those related to makeup and fingernails [8, 18, 21, 31, 66]. In the field of makeup, makeup transfer [18, 21, 31] aims to transfer the makeup style of a reference image to the source image. In nail-related research, NailNet [8] is a segmentation method that segments the fingernail and the lunula, which can be applied to virtual nail art. Moreover, Yarimbiyik et al. [66] propose a model that recommends the nail shapes that suit the shape of the fingers. Several studies have proposed fashion-centric vision-language foundation models [15, 16, 52, 74] that tackle downstream multimodal tasks, such as cross-modal retrieval [9, 49], category recognition [49, 59], fashion image captioning [49, 65], and text-guided image retrieval [62].

Contrastive learning. Contrastive learning is widely adopted in vision-language foundation models (e.g., [29, 48]). CLIP employs InfoNCE [44] as its contrastive loss function. However, training with InfoNCE assumes a strict one-to-one correspondence between a single text and a single image as a positive example. Consequently, samples that are semantically aligned but not explicitly labeled as positive pairs are treated as negatives, introducing noise into the learning process [61]. To address this issue, various studies have relaxed the one-to-one labeling constraint in InfoNCE-based training (e.g., [10, 34, 61, 67]). Existing approaches typically handle such ambiguous pairs by (i) leaving them on the negative side with reduced weight, (ii) incorporating them as additional positives via soft targets or multi-positive alignment, or (iii) masking or excluding them from training. Because unlabeled positives are plausible targets yet not necessarily exact matches, these strategies still have limitations: (i) can still repel near-matches as negatives, (ii) can collapse graded partial matches by pulling them toward exact positives or by continuously adjusting similarity toward a soft target, and (iii) discards useful supervision. For example, ReCo [34] mitigates this issue by relaxing the constraints on negative pairs, ignoring those with similarity scores below zero, but it does not explicitly handle unlabeled positives.

Benchmarks on fashion AI. As mentioned above, there are several benchmarks for multimodal retrieval in fashion AI [12, 49, 62]. In the field of image synthesis, research has been conducted on tasks such as virtual try-on [4, 13, 43] and makeup transfer [18, 21, 31]. For instance, VITON-HD [4] serves as a benchmark for evaluating high-resolution virtual try-on methods, focusing on garment detail preservation, misalignment handling, and realistic synthesis. In addition, the Makeup Transfer dataset [31] is utilized to evaluate the effectiveness and robustness of makeup transfer techniques across various makeup styles and lighting conditions.

Several benchmarks focus on nail-related research [6, 8, 41, 51]. These studies primarily focus on segmentation tasks and are applied in various domains, including assisting disease assessment in medical diagnostics, forensic personal identification, and the beauty industry for virtual nail art applications [8].

Comparison with existing benchmarks. The NAIL-STAR benchmark constructed in this study differs from existing benchmarks in the following aspects. First, existing benchmarks in the fashion AI field [12, 49, 62] often focus on the multimodal retrieval of pre-manufactured products. In contrast, the NAIL-STAR benchmark incorporates nail design images that are largely composed of both: (i) a painted part allowing for creative flexibility and (ii) a decorative part that can be modified only by selecting and arranging pre-manufactured embellishments. Thus, the NAIL-STAR benchmark requires the models to handle designs that are more diverse than those of existing fashion benchmarks. Moreover, most existing benchmarks feature relatively simple and superficial comprehension of multimodal contents. Most nail image benchmarks do not include a broad spectrum of realistic designs, nor do they provide multi-layered descriptions of the underlying intent of the designs.

$\boldsymbol{x}_{\text{txt}}$: "The design features clear nails decorated with hand-painted floral patterns in shades of blue, light blue, and yellow, accented with delicate butterfly ornaments."

Figure 5. Examples of c_{ij} estimated by the MLLM in CRAM. Positive, unlabeled positive, and negative labels are framed in green, yellow, and red, respectively, with unlabeled positives identified based on c_{ij}.

B. Prompt Examples

IPFM. The prompt used in Intent-Palette Fusion Module (IPFM) to obtain $\boldsymbol{x}_{\text{MDD}}$ from $\boldsymbol{x}_{\text{txt}}$ was as follows:

```
Extract the design, shape, theme, and impression
details from the following nail design request.
Only include details that are directly stated or
clearly described in the text. Do not infer or
mention missing parts.

- Design: Visual details like colors, patterns,
textures, or nail accessories.
- Shape: Nail shape or length.
- Theme: The theme, motif, or symbol that
represents the design.
- Impression: The impression conveyed by the
design or theme.

If any part (design, shape, theme, impression) is
 not explicitly mentioned, exclude it from the
output. There is no need to state that it is
missing.

Use this format for the output. Exclude any
categories that are not explicitly mentioned:
The design of these nails is ....
The shape of these nails is ....
The theme of these nails is ....
The impression of these nails is ....

nail design request: {dense intent description}
```

We replaced the above-mentioned placeholder in `{dense intent description}` with $\boldsymbol{x}_{\text{txt}}$. Similarly, the prompt template to obtain $\boldsymbol{x}_{\text{NNP}}$ from $\boldsymbol{x}_{\text{txt}}$ was defined as follows:

```
The following text is a request for a nail design
 to a nail artist. From this request sentence,
create a noun phrase that describes the design,
theme, or impression of the nail designs. The
sentence can be long, but it should retain all
the information and be presented as a single noun
 phrase. Respond with only the noun phrase.

nail design request: {dense intent description}
```

As an example, suppose $\boldsymbol{x}_{\text{txt}}$ was as follows: *"I'd like a design that primarily features purple, inspired by a wizard, to create a mysterious atmosphere. Please add decorations resembling a ruby on the ring finger. I want long nails with dark marble patterns to enhance the mysterious vibe."* For this example, the following output for $\boldsymbol{x}_{\text{MDD}}$ was obtained:

```
The design of these nails is primarily purple
with dark marble patterns and decorations
resembling a ruby on the ring finger. The shape
of these nails is long. The theme of these nails
is inspired by a wizard. The impression of these
nails is mysterious.
```

Similarly, $\boldsymbol{x}_{\text{NNP}}$ was obtained as follows:

```
A design primarily featuring purple inspired by a
 wizard with a mysterious atmosphere, long nails
with dark marble patterns, and ruby-like
decorations on the ring finger.
```

VDFM. The following prompt was used in Visual Design Fusion Module (VDFM) to generate explanations based on $\boldsymbol{x}_{\text{img}}$ using MLLMs:

```
Please describe the nail designs in the image by
highlighting the important features, such as
colors, patterns, decorations, textures, nail
shapes, nail lengths, motifs, or impressions.
There's no need to cover everything--focus only
on the key distinctive elements.
```

Suppose the target nail design image in Fig. 4 (a) is given as input. The following outputs were generated depending on the MLLM used. GPT-4o generated the following output:

```
The nails are long with a coffin shape, featuring
 a rainbow gradient of bright colors including
pink, blue, green, orange, and yellow. Notably,
each hand has a focal point decoration: a large,
multicolored 3D flower with a smiling face at the
 center, adding a playful and bold touch. The
designs are glossy, which enhances their vibrant
appearance.
```

In contrast, Qwen2-VL generated:

```
The nails in the image feature a vibrant and
playful design with the following key elements:\n
\n1. **Colors**: The nails are painted in a
gradient of bright, neon colors, including pink,
blue, green, yellow, and orange. The gradient
transitions smoothly from one color to the next.\
n\n2. **Patterns**: The nails have a striped
pattern with alternating bands of the same colors
 as the gradient. This creates a cohesive and
visually appealing design.\n\n3. **Decorations**:
 Each nail is adorned with a colorful, three-
dimensional flower decoration. The flowers are
made of small, round beads in various colors,
including red, yellow, green, blue, and purple.
Each flower has a smiling face in the center,
adding a cheerful and whimsical touch.\n\n4. **
Textures**: The nails appear to have a glossy
finish, giving them a shiny and reflective look.\
n\n5. **Nail Shapes**: The nails are long and
coffin-shaped, which is a popular style known as
\"stiletto\" nails.\n\n6. **Lengths**: The nails
are quite long, extending beyond the natural nail
 bed, which is typical for stiletto nails.\n\
nOverall, the nails combine bright colors,
playful patterns, and cute decorations to create
a fun and eye-catching design.
```

CRAM. We estimated c_{ij} in Confidence-based Relaxed Alignment Module (CRAM) using MLLMs. The MLLM output was either "0," "5," or "10," which were scaled to 0.0, 0.5, or 1.0, respectively, and used as c_{ij}. The following prompt was used to estimate c_{ij}, which represents the degree to which $\boldsymbol{x}_{\text{img}}^{(j)}$ can be considered as an unlabeled positive for $\boldsymbol{x}_{\text{txt}}^{(i)}$:

```
Your task is to evaluate how well the nails shown
 in the image match the provided description.
Analyze these aspects: colors, patterns, textures
, decorations, motifs, nail lengths, nail shapes,
 and overall impressions. The image is annotated
as {candidate noun phrase}. Can it also be
described as {gt noun phrase}?
Provide a score of 0, 5, or 10 based on the
following criteria:
10: A flawless match with no visible differences
or inconsistencies in any aspect.
 Use this score only if the image perfectly
aligns with {gt noun phrase}, leaving no room for
 doubt.
5: The image is very close to being perfectly
described as {gt noun phrase}, with only minor
and barely noticeable differences
 limited to slight variations in color or nail
shape. Differences in illustrations, motifs, or
other prominent aspects should result in a score
of 0.
0: Use this score even if most aspects are
similar, but there is at least one significant
and noticeable difference
 that prevents the image from being described as
{gt noun phrase}.
State the score (0, 5, or 10) first, followed by
a concise reason.
```

The above-mentioned placeholders `{candidate noun phrase}` and `{gt noun phrase}` were replaced with $\boldsymbol{x}_{\text{NNP}}^{(j)}$ and $\boldsymbol{x}_{\text{NNP}}^{(i)}$, respectively. Fig. 5 shows the resulting scores. Images that are closely aligned with $\boldsymbol{x}_{\text{txt}}$ are assigned higher scores, whereas those that only partially match $\boldsymbol{x}_{\text{txt}}$ receive lower scores. In the experiment, MLLM outputs of "10" or "5" for $\boldsymbol{x}_{\text{img}}^{(j)}$ were classified as unlabeled positives for $\boldsymbol{x}_{\text{txt}}^{(i)}$ because we set $\theta = 0.5$.

C. Accelerating Confidence Score Estimation

In CRAM, we introduce a preprocessing method that speeds up confidence score estimation by filtering candidate pairs according to similarity. Estimating the scores for all possible pairs in the training set requires performing (number of descriptions) $\times$ (number of images) estimations, resulting in a significant computational cost. Therefore, it is desirable to perform the estimation process on only a subset of samples that have a high probability of being unlabeled positives. To achieve this, we first obtain the language feature and the visual feature from $\boldsymbol{x}_{\text{txt}}^{(i)}$ and $\boldsymbol{x}_{\text{img}}^{(j)}$, respectively, using a vision–language model (e.g., CLIP [48], BEiT-3 [60]). We then compute their similarities using cosine similarity [48]. Because an image $\boldsymbol{x}_{\text{img}}^{(j)}$ with high similarity to $\boldsymbol{x}_{\text{txt}}^{(i)}$ is likely an unlabeled positive, we select the top N_{cand} images $\left\{\boldsymbol{x}_{\text{img}}^{(k)}\right\}$ (where k represents each element in the index set of these top N_{cand} images) with the highest similarity scores as candidate unlabeled positives.

D. NAIL-STAR Benchmark

D.1. Image Filtering

The collected nail images were pre-processed using the following filtering steps. (1) Images that showed toenail designs and nail tips were excluded because this study focuses solely on fingernail designs. For this purpose, we employed CLIP-based filtering, an established method [7, 23] for image selection in dataset construction, leveraging the similarity between each image and the text provided below. Specifically, images were removed if they met either of the following criteria: (a) the similarity to "a photo of fingers" or "a photo of nails" fell below a predetermined threshold, or (b) the similarity to "a photo of hands" was lower than that to "a photo of feet."

(2) Images in which the nail area was disproportionately small relative to the overall image were excluded. This is because it is difficult to recognize nail designs in images where the nail area appears small because of the distance from the camera, or where much of the nail is obscured by the angle of the hand. (3) Images where two or fewer nails were visible were excluded as this made it difficult to grasp the overall design. (4) Images featuring 11 or more nails were removed because they were presumed to include more

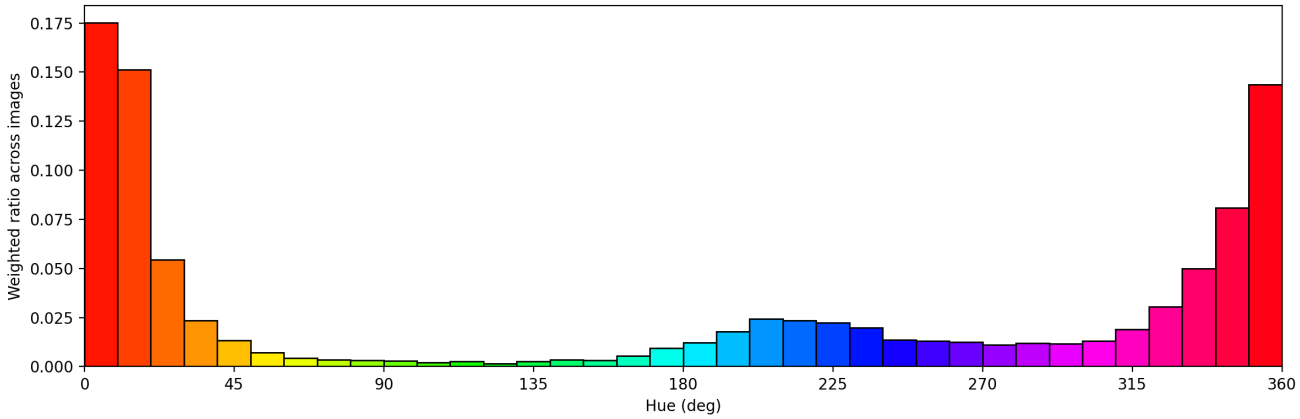

Figure 6. Hue histogram of nail design images. We compute the distribution of hue values (HSV space, 0–360°) using all pixels within segmented nail regions in the NAIL-STAR benchmark. To visualize hue independent of saturation and value, the color of each bar corresponds to the bin center with fixed parameters ($S = 1, V = 1$). The vertical axis indicates the pixel proportion relative to the total segmented area across the dataset.

than one person's hands, potentially containing a mixture of different nail designs. As a preprocessing step for the filtering, nail segmentation masks were generated for the images using a segmentation model for fingernails [42]. Then, criteria (2) to (4) were applied based on the area and number of nails identified.

(5) To eliminate duplicate images, the Hamming distance between the hash values of each image was calculated, and identical or extremely similar images were filtered out. (6) A manual inspection was conducted to exclude any remaining images deemed unsuitable for this study, such as AI-generated images or those with extremely poor lighting that hindered the recognition of nail designs.

Fig. 6 shows the distribution of hue values across all segmented nail regions in the NAIL-STAR benchmark. We extracted pixel values in the HSV color space and constructed a histogram discretized over the range 0°–360°. The high proportion of warm colors is attributable not only to a bias toward warm-colored nail designs, but also to the warm hue of the underlying nail plate visible through clear gel.

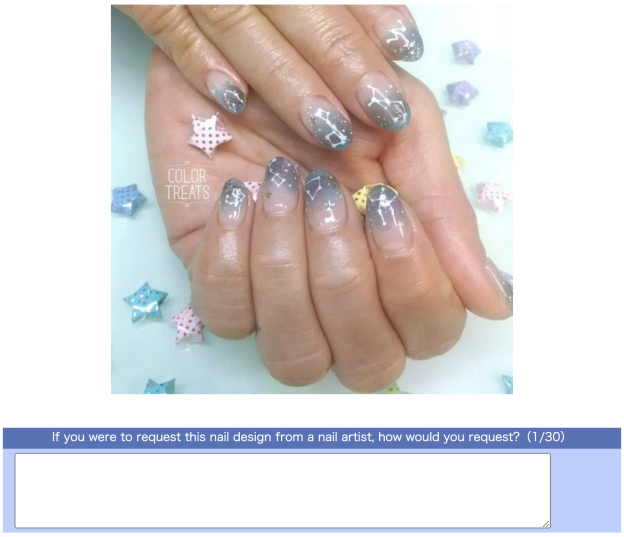

Figure 7. Description annotation interface. The annotators were requested to input a dense intent description for the displayed nail design image into the text box below the image.

D.2. Annotation of Descriptions

We annotated dense intent descriptions for nail design images as described below. Fig. 7 shows the annotation interface. The annotators were presented with a series of nail design images, one at a time, and were instructed to respond to the following question: *"If you were to request this nail design from a nail artist, how would you request it?"* They were instructed to provide a visual description of each nail design, encompassing elements such as color, pattern, texture, and nail embellishments. Moreover, they were asked to include, where relevant, any themes suggested by the visual characteristics and the overall impression conveyed by the design. To avoid simple image captions, the annotators were instead instructed to imagine a scenario in which they were describing a nail design to a nail artist and to express their intent in a detailed and multi-layered manner.

We implemented a multi-step quality control process for the reliability of the crowdsourced annotations. First, we randomly sampled and manually inspected five annotations from each annotator. All data from annotators with systematic errors were excluded. Second, we employed the Polos score [56], an automatic evaluation metric for image captions, to identify potentially low-quality annotations. All samples with a Polos score below 0.2 were manually reviewed, and erroneous annotations were removed. Finally, every sample in the test set was manually checked.

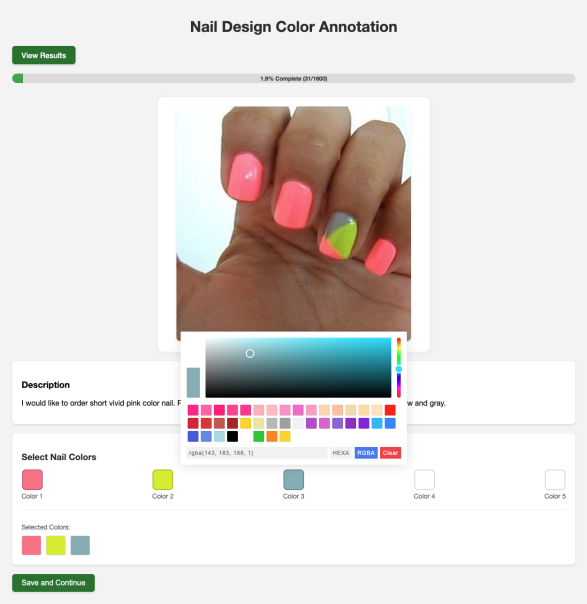

Figure 8. Palette query annotation interface. The annotators were requested to specify a palette query for the displayed nail design by using a color picker, selecting any number of colors (0–5).

Data collection was facilitated through the use of SoSci Survey. All descriptions were required to be in English and at least 10 words. A total of 10,625 descriptions were collected from 208 annotators, with an average of 51.1 descriptions per annotator. Fig. 9 shows the word frequency distribution of the annotations.

D.3. Annotation of Palette Queries

We automatically annotated palette queries for the training and validation sets of the NAIL-STAR and Marqo Fashion200K benchmarks as follows. First, we generated segmentation masks from X_{img} using a fingernail segmentation model [42] for the NAIL-STAR benchmark and a segmentation model [63] fine-tuned on a fashion dataset for the Marqo Fashion200K. Within each segmented region, we applied SLIC to partition the area into N_{SLIC} superpixels, where $N_{\text{SLIC}} \geq 1000$. For each superpixel, we computed the mean RGB value and pixel count, defining this mean as the superpixel's representative color. These colors were subsequently converted to the CIELAB color space to enable perceptually meaningful comparisons.

Next, we computed a pairwise distance matrix between representative colors in the CIELAB space using the CIEDE2000 color difference ΔE_{00} and performed agglomerative clustering with average linkage. Clusters were iteratively merged until the average inter-cluster distance exceeded a threshold θ_{cluster}. The representative color of each cluster was defined as the color of the superpixel that minimized the weighted sum of ΔE_{00} distances to all other superpixels in the same cluster, and its CIELAB value was converted back to RGB.

Finally, we sorted the clusters by area ratio in descending order to select the palette query. Iterating from the largest to the smallest cluster, we retained a cluster as a color of the palette query only if its individual area ratio was at least θ_{min} and the cumulative area ratio of selected clusters did not exceed θ_{cum}.

In contrast, we manually annotated palette queries for the test sets of both benchmarks as follows. As illustrated in Fig. 8, annotators were presented with both the target image and its corresponding description, utilizing an interface that mimicked a typical e-commerce search UI with a color picker. Annotators were allowed to select between zero and five colors for each sample. Rather than exhaustively specifying all colors present in the image, they were instructed to choose only the key colors they would input if they were actually searching for that design. For efficiency, we provided a set of representative colors as initial candidates. However, annotators were not restricted to these presets and could freely adjust the color picker to any arbitrary color before finalizing the query.

E. Implementation Details

We trained our model on a GeForce RTX 4090 with 24 GB of VRAM and an Intel Core i9-13900KF with 64 GB of RAM. The training time for the proposed model was approximately 20 minutes. All features and generated outputs from LLMs and MLLMs required for training were obtained offline prior to the training process, because these models remain frozen and do not need to be recomputed during training. The average total computational time per sample for feature extraction and generation by LLMs and MLLMs was approximately 4.2 seconds. At inference time, the similarity between a single description and 1,600 nail design images was computed in approximately 0.8 seconds, where the features from $\boldsymbol{x}_{\text{txt}}$ were obtained online, while those from $\boldsymbol{x}_{\text{img}}$ were pre-extracted. In practical text-to-image retrieval, the candidate pool of images is predetermined. Therefore, it is feasible to obtain and store the visual features in advance, thereby significantly reducing the computational burden during inference. Moreover, this latency of 0.8 seconds is considered acceptable, as users typically do not perceive delays under 1 second as problematic [2].

We used the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and a fixed learning rate of 1×10^{-5}. We set the batch size to 64 and trained the model for 40 epochs. The number of candidate images was set to $N_{\text{cand}} = 30$. The weights of the loss functions were set to $\lambda_{\text{UP}} = 0.7$, $\lambda_{\text{N}} = 0.7$. The language representations $(\boldsymbol{l}_{\text{txt}}, \boldsymbol{l}_{\text{MDD}}, \boldsymbol{l}_{\text{NNP}})$, the palette representation $\boldsymbol{p}$, and the language–palette representation $\boldsymbol{l}_+$ described in Section 3.3, as well as the visual representations $\boldsymbol{v}_{\text{s}}^{(i)}$, $\boldsymbol{v}_{\text{a}}^{(i)}$, $\boldsymbol{v}_{\text{n}}^{(i)}$, and $\boldsymbol{v}$ detailed in Section 3.4, are all 1024-dimensional. In the CRAM, the MLLM output was either "0," "5," or "10," which were scaled to 0.0, 0.5, or 1.0,

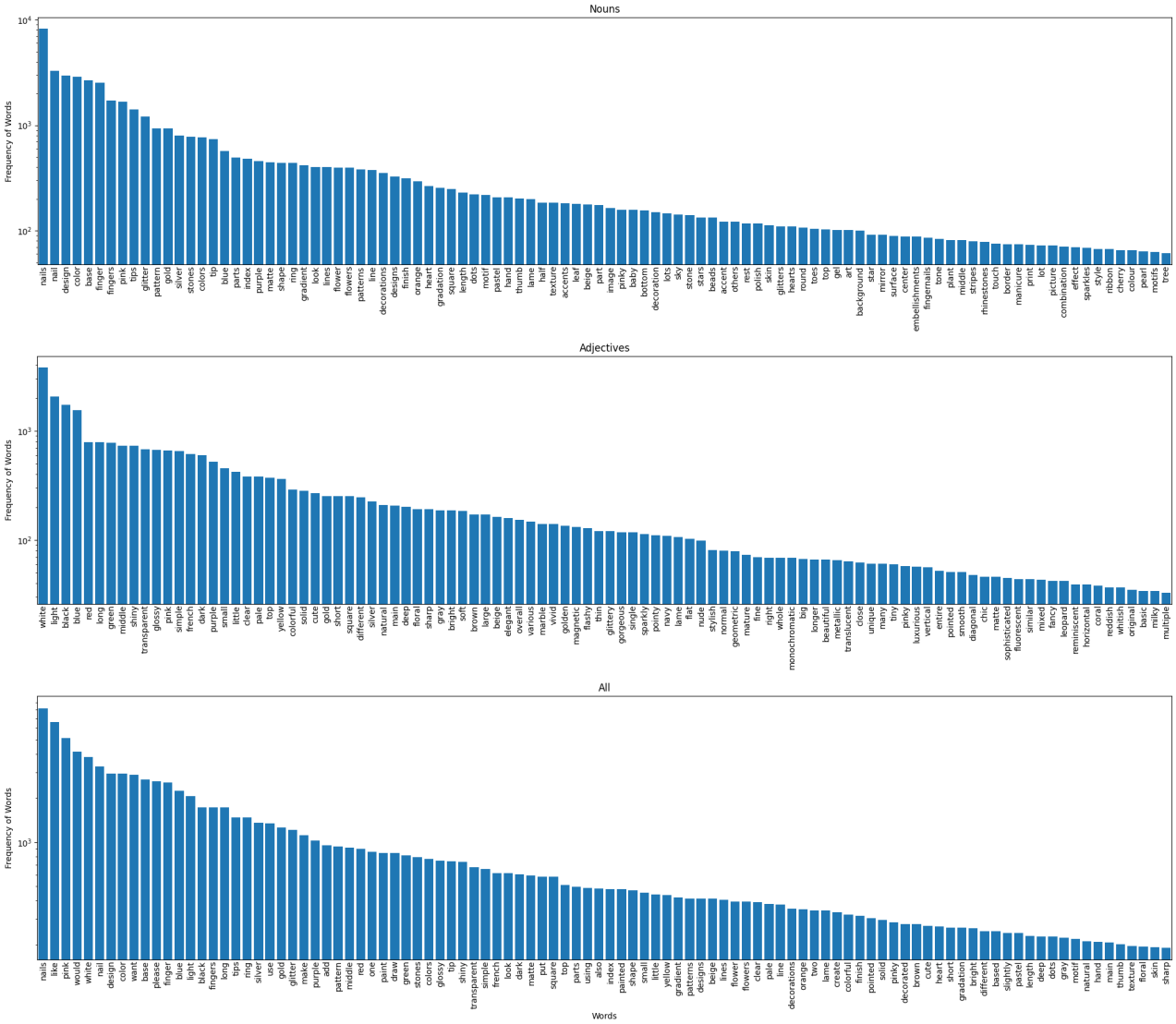

Figure 9. The word frequency distribution of the NAIL-STAR benchmark. The frequency distributions of the 100 most common nouns, adjectives, and all words are presented. Singular and plural forms of nouns, comparative and superlative forms of adjectives, and present and past participles of verbs were counted independently. Stopwords were removed.

respectively, and used as c_{ij}. Our model had approximately 4.3×10^7 trainable parameters and 1.2×10^9 multiply-add operations. We calculated the recall@1 on the validation set for each epoch. For the evaluation on the test set, we used the model with the maximum recall@1 on the validation set.

F. Discussion on Token Length Limitations

In our experiments, all models were fine-tuned on the NAIL-STAR and Marqo Fashion200K benchmarks, where descriptions exceeding the maximum token length of each text encoder were truncated. However, the token length limitations of text encoders did not significantly affect the performance of the methods in the comparisons. Among all baselines, SigLIP [69] has the shortest maximum input length, limited to 64 tokens. Our method employed BEiT-3 [60] and SigLIP as text encoders. To ensure a fair comparison, the BEiT-3 text encoder was consistently configured with a maximum input length of 64 tokens in both the baseline and our proposed method. The other baseline methods were evaluated using the default maximum token lengths defined in their official implementations (e.g., 77 tokens for CLIP [48]). Therefore, the token length limitations did not provide any advantage to our method. Furthermore, in the NAIL-STAR benchmark, only 0.42% of the descriptions exceed 64 tokens, suggesting that the impact of information loss due to truncation on the evaluation results was minimal.

Method	MRR↑ [%]	R@1↑ [%]	R@10↑ [%]
CLIP [48]	12.4	6.5	23.7
FashionViL [14]	2.3	0.6	4.8
FAME-ViL [15]	20.2	11.2	39.3
BEiT-3 [60]	34.9	23.5	57.9
BLIP-2 [30]	14.4	7.5	28.0
SigLIP [69]	31.4	21.3	51.6
Alpha-CLIP [54]	19.6	12.1	34.3
Long-CLIP [70]	10.6	5.6	19.7
MM-Embed [32]	15.3	9.1	27.8
LamRA [36]	15.0	9.9	28.6

Table 5. Quantitative results of baseline methods on the NAIL-STAR benchmark in the zero-shot settings. The scores are reported based on a single trial.

G. Evaluation Metrics

We used the mean reciprocal rank (MRR) and recall@K($K = 1, 10$) as evaluation metrics, with recall@1 as the primary metric. The MRR is defined as follows:

$$\text{MRR} = \frac{1}{N_{\text{txt}}} \sum_{i=1}^{N_{\text{txt}}} \frac{1}{r_1^{(i)}}, \tag{8}$$

where N_{txt} and $r_1^{(i)}$ denote the number of descriptions and rank of the target nail design image for the i-th description, respectively. Recall@K is defined as follows:

$$\text{Recall@}K = \frac{1}{N_{\text{txt}}} \sum_{i=1}^{N_{\text{txt}}} \frac{|A_i \cap B_i|}{|A_i|}, \tag{9}$$

where A_i and B_i denote the set of nail design images to be retrieved for the i-th description and the set of top-K images for the i-th description, respectively.

H. Additional Quantitative Results

H.1. Zero-Shot Evaluation

Table 5 shows the results of the zero-shot evaluation of the baseline models on the NAIL-STAR benchmark. The scores are reported based on a single trial, since these provide consistent results across multiple trials. Even for the best-performing method BEiT-3, the recall@1 score remains at only 23.5%. This result highlights the difficulty of this task.

H.2. Effect of Query Length on Performance

To analyze how performance varies with query length, we conducted additional evaluations. We divided the test set into subsets according to the token length of the descriptions and evaluated NaiLIA in the description-only setting, where $\boldsymbol{x}_{\text{pal}}$ is not provided. Table 6 shows the distribution of token lengths and the evaluation results for each subset. The results indicate that MRR and recall@1 scores increase as the token length grows. These findings demonstrate that NaiLIA is robust across all token lengths and is particularly effective for longer descriptions containing dense intent.

Token Length	#	MRR↑ [%]	R@1↑ [%]	R@10↑ [%]
0–19	781	57.9	45.3	82.5
20–29	560	60.0	48.0	81.4
30–39	187	68.4	58.3	89.3
40–69	72	78.6	69.4	94.4
Total	1600	60.8	48.9	83.4

Table 6. Retrieval performance of NaiLIA across subsets grouped by description token length on the NAIL-STAR benchmark, evaluated in the description-only setting without $\boldsymbol{x}_{\text{pal}}$. The scores are reported based on a single trial.

Colors	#	SigLIP	NaiLIA
0	1600	47.5	49.5
1	386	48.6	50.9
2	738	53.9	57.5
3	393	56.3	60.1
≥4	83	62.8	66.7

Table 7. Recall@1 of NaiLIA and SigLIP across test subsets grouped by palette cardinality on the NAIL-STAR benchmark. The scores are reported based on a single trial.

H.3. Effect of Query Length on Performance

To analyze the impact of palette cardinality, we conducted an additional evaluation by partitioning the test set into subsets according to the number of colors in the palette query, while keeping the full test set as the retrieval pool. Table 7 shows Recall@1 for the best baseline (SigLIP) and NaiLIA under the same palette conditioning. NaiLIA consistently outperforms SigLIP regardless of palette size, indicating robustness under varying levels of user input complexity. For both methods, performance improves as more colors are provided, since higher-cardinality palettes offer more informative inputs. Moreover, the performance gap between SigLIP and NaiLIA becomes larger with two or more colors, where the model should capture how each color relates to other design attributes (e.g., patterns, decorations, and finger placement). The widening margin therefore suggests that NaiLIA understands the semantics of the dense intent descriptions beyond simple color matching.

H.4. Ablation Study of VDFM

To demonstrate the generality of VDFM, we conducted an ablation study on the Marqo Fashion200K. Table 8 shows that recall@1 drops by 1.2, 46.0, and 4.8 points for Models (b), (c), and (d), respectively, compared with Model (a). This suggests that the VDFM is impactful across domains and not limited to a specific setting.

Model	MRR↑ [%]	R@1↑ [%]	R@10↑ [%]
(a) NaiLIA (full)	82.5	74.5	96.6
(b) w/o $\boldsymbol{v}_\mathrm{s}$	81.6	73.3	95.7
(c) w/o $\boldsymbol{v}_\mathrm{a}$	41.1	28.5	65.7
(d) w/o $\boldsymbol{v}_\mathrm{n}$	78.8	69.7	95.1

Table 8. Ablation study of VDFM on the Marqo Fashion200K. The scores are reported based on a single trial.

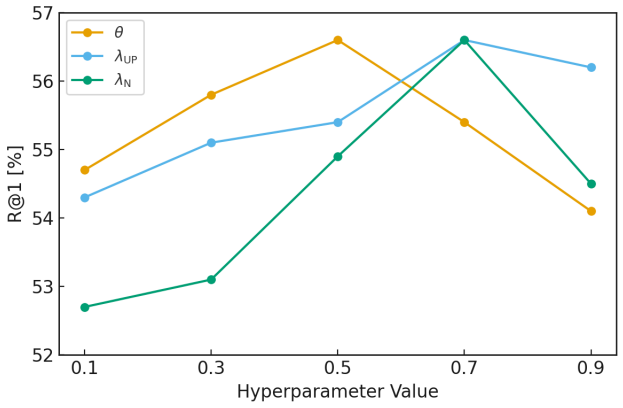

Figure 10. Sensitivity analysis of the CRC loss with respect to the confidence threshold θ and the balance terms λ_UP and λ_N. For each curve, we varied one hyperparameter over $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ while keeping the other two fixed at their default values ($\theta = 0.5$, $\lambda_\mathrm{UP} = 0.7$, $\lambda_\mathrm{N} = 0.7$). The scores are reported based on a single trial.

H.5. Sensitivity Analysis

We conducted a sensitivity analysis to examine how each hyperparameter in the CRC loss influences retrieval performance, as illustrated in Fig. 10. In this analysis, we varied one hyperparameter over $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ while keeping the remaining two fixed at their default values ($\theta = 0.5$, $\lambda_\mathrm{UP} = 0.7$, $\lambda_\mathrm{N} = 0.7$). The scores are reported based on a single trial. The results demonstrate the robustness of the CRC loss to variations in hyperparameters. Similar trends were also observed across other parameter combinations.

H.6. MLLM-Based Retrieval

We employed MLLMs to estimate c_{ij} in CRAM because MLLMs can approximate similarity scores between text and image pairs. However, retrieval based solely on similarity scores predicted by MLLMs is impractical due to their high computational cost and limited performance. In practice, computing the similarity between a single description and 1,600 nail design images took an average of 18.7 minutes. When performing inference on 10 descriptions, the recall@1 was limited to 20.0% (with the target nail design image ranked 28th on average). The experiment was conducted using Qwen2-VL [58] as the MLLM, with a batch size of 128 for batch processing. The MLLM was prompted to output an integer similarity score from 0 to 9 for each image, and the images were ranked in descending order of these scores. In cases where multiple images were assigned the same score, their order was randomized within the tied rank. These results highlight the practical limitations of using MLLMs directly for retrieval.

I. Error Analysis

To investigate the limitations of the proposed method, we analyzed cases where the method did not perform as expected. We analyzed the 100 failure cases with the lowest ranks for the target nail design images. Table 9 shows the categorizations of the failure modes for both the description-only and full settings of NaiLIA on the NAIL-STAR benchmark. The failure modes could be grouped into the following eight categories:

- Comprehension errors in patterns and decorative elements: This category encompasses cases where the model incorrectly interpreted patterns or decorative elements in the nail design. This includes cases where these elements were relatively small compared with the overall image or had colors similar to the background, making them difficult to recognize.
- Comprehension errors in design themes: This category covers cases where the model failed to comprehend the themes of the nail design, such as those inspired by specific objects or concepts.
- Comprehension errors in color-related expressions: This refers to errors where the model wrongly interpreted color-related expressions. For example, there were cases where nail designs with separate pink and gray colors were ranked higher than those with a blended hue, which was misaligned with user intent, even when the description specified "a color that's between pink and gray."
- Comprehension errors in nail shape and decoration placement: This category includes cases where the nail length, shape, or decoration placement was misinterpreted by the model. For example, there were cases where nail designs with glitter applied at the base rather than the tips were included, even when the description explicitly specified "glitter on the tips."
- Comprehension errors in the correspondence between fingers and design: This category refers to errors in which the model failed to comprehend the correspondence between the finger and the design. Specifically, this includes cases where images in which the painting or decoration of the nail on a different finger was incorrectly ranked higher than the target nail design image.
- Ambiguity in color description: This error arises from the ambiguity inherent in color description. For example,

Errors Type	Description	#Error (desc-only)	#Error (full)
Comprehension	Comprehension errors in patterns and decorative elements	**19**	**20**
	Comprehension errors in design themes	17	19
	Comprehension errors in color-related expressions	13	6
	Comprehension errors in nail shape and decoration placement	11	11
	Comprehension errors in the correspondence between fingers and design	5	7
Ambiguity	Ambiguity in color description	18	9
	Ambiguity in impression description	9	18
	Ambiguous descriptions unrelated to color or impression	8	10
	Total	100	100

Table 9. Number of failure cases in each category for NaiLIA in the description-only and full settings on the NAIL-STAR benchmark. The largest count in each setting is highlighted in **bold**.

given a description specifying "pink," the model preferentially retrieved nail designs with a reddish-pink hue, even if the user wanted a yellowish one.

- Ambiguity in impression description: This category comprises cases where the given description was ambiguous in terms of impression-related expressions, such as "flashy" or "mysterious."
- Ambiguous descriptions unrelated to color or impression: This refers to cases where numerous unlabeled positive samples dominated the top rankings.

J. Discussion on Computational Trade-offs

Our multi-representation fusion is an intentional design choice to capture user intent across diverse levels of abstraction, ranging from concrete visual details to high-level concepts. However, when computational resources are limited, the architecture can be simplified by removing selected branches, highlighting the modular flexibility of the proposed method. For example, the ablation study in Table 2 suggests that the unimodal encoder branch $\boldsymbol{v}_s$ is a natural candidate for omission, and its removal results in a 1.3-point drop in Recall@1. In contrast, removing the MLLM-based branch $\boldsymbol{v}_n$ results in a 2.4-point drop. In practice, a lighter variant of $\boldsymbol{v}_n$ can be considered, and rapid progress in open, lightweight MLLMs may make it increasingly computationally cost-effective.

K. Discussion on LLM Hallucinations

We investigated how potential LLM hallucinations or parsing errors might propagate to retrieval quality. Specifically, for the 100 failure cases analyzed in Section I, we manually examined whether any hallucinations were present in the multi-layered design descriptions and normalized noun phrases corresponding to each description. As a result, none of the 100 failure cases analyzed involved hallucinations caused by the LLM. This can be attributed to the fact that the prompts explicitly instructed the LLM to use only information present in the original text, thereby reducing the likelihood of hallucination. Although hallucinations may exist outside these cases, the integration of three types of features in IPFM appears to have prevented such hallucinations from resulting in critical errors.

L. Additional Qualitative Results

Fig. 11 presents additional qualitative results of NaiLIA in the description-only setting, where only $\boldsymbol{x}_{\text{txt}}$ is provided. Fig. 12 and Fig. 13 show results of NaiLIA in the setting where both $\boldsymbol{x}_{\text{txt}}$ and $\boldsymbol{x}_{\text{pal}}$ are provided. The top-5 retrieved images are shown. Positives and unlabeled positives are enclosed in green and yellow frames, respectively.

M. URLs of Images in This Paper

The images included in this paper are available at the URLs listed in Table 10 and Table 11.

Figure 11. Additional qualitative results of NaiLIA in the **description-only** setting, where only x_{txt} is provided. The top-5 retrieved images are shown. Positives and unlabeled positives are enclosed in green and yellow frames, respectively.

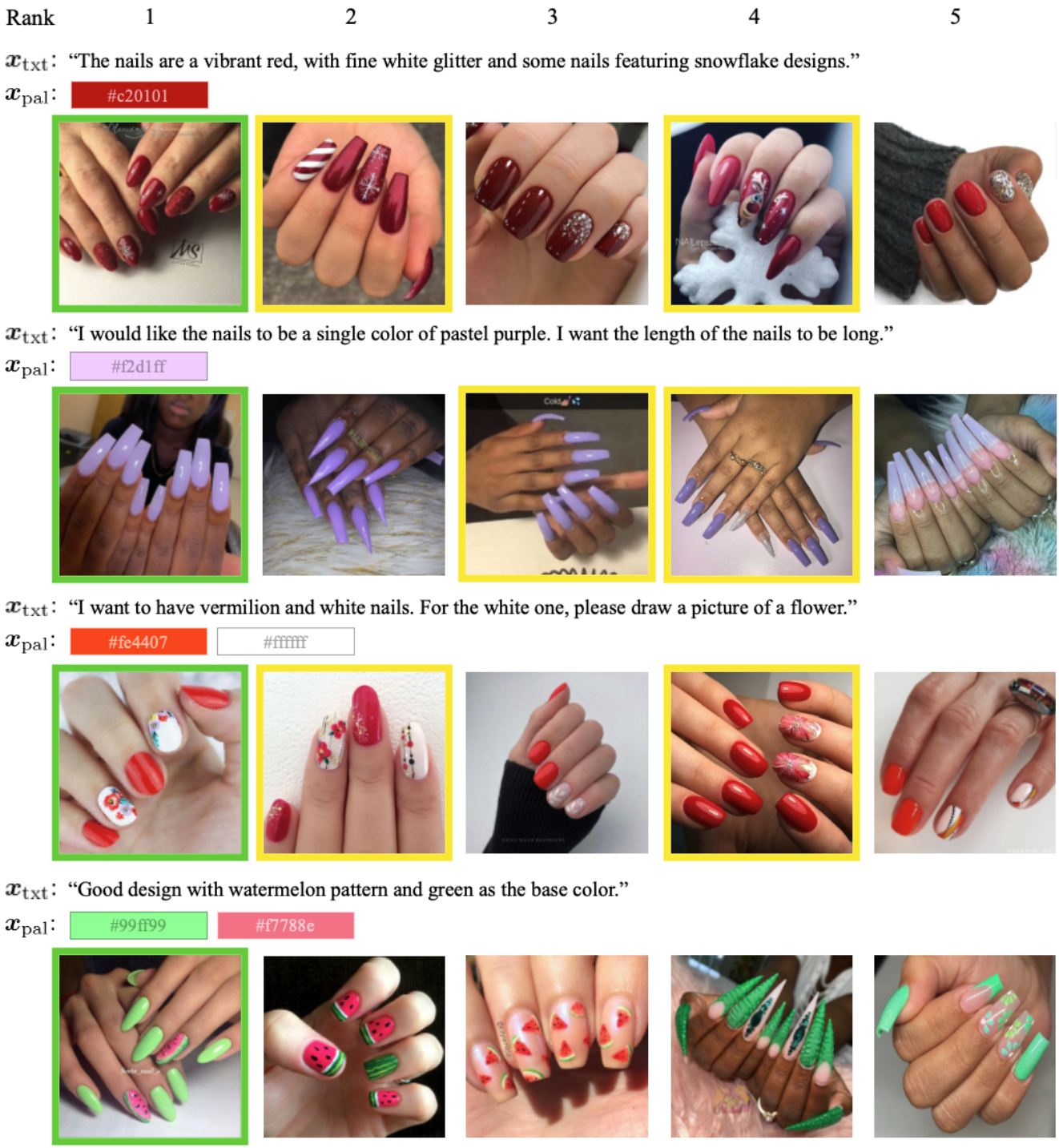

Figure 12. Additional qualitative results of NaiLIA in the setting where **both** x_{txt} and x_{pal} are provided. The top-5 retrieved images are shown. Positives and unlabeled positives are enclosed in green and yellow frames, respectively.

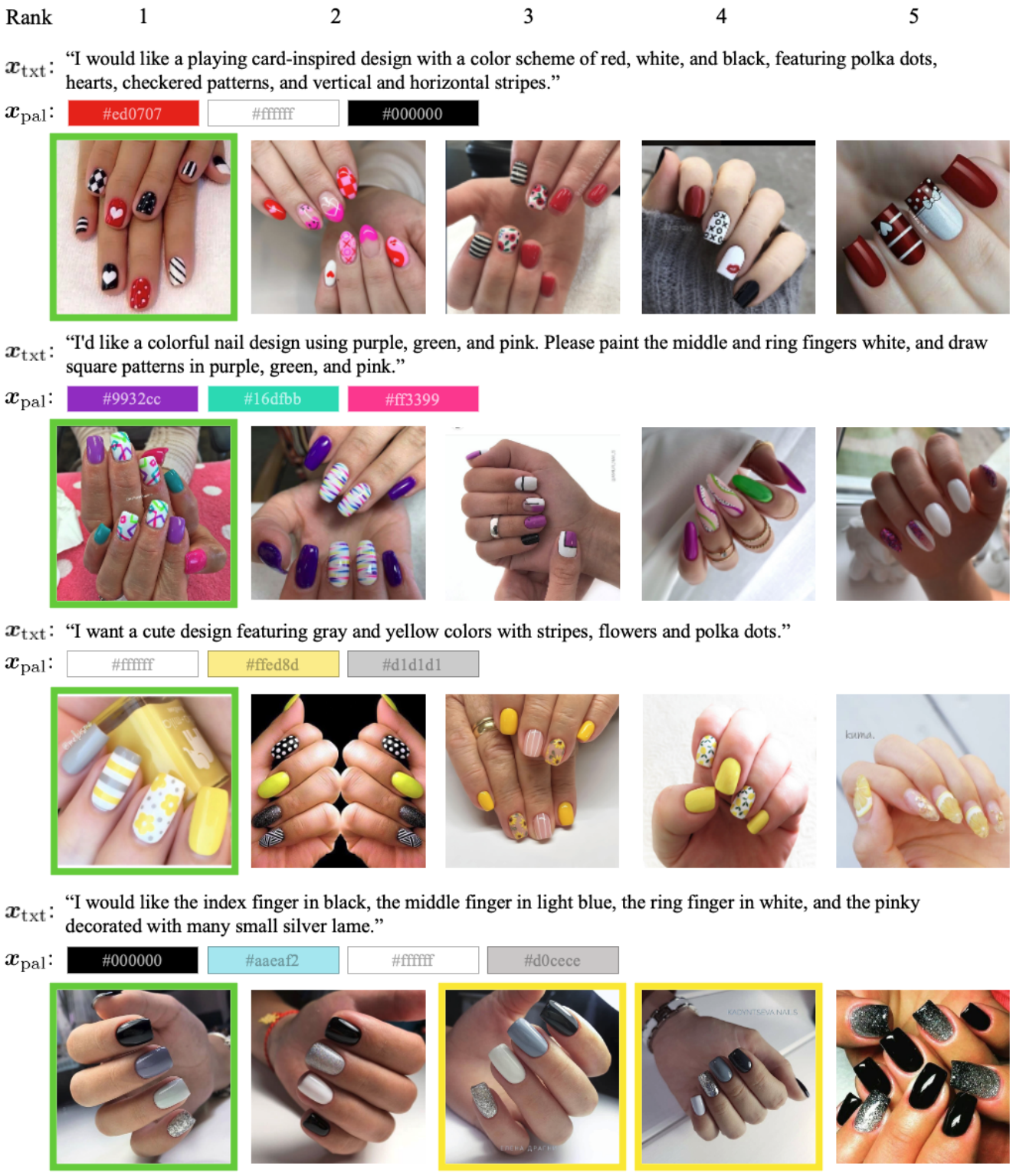

Figure 13. Additional qualitative results of NaiLIA in the setting where **both** x_{txt} and x_{pal} are provided. The top-5 retrieved images are shown. Positives and unlabeled positives are enclosed in green and yellow frames, respectively.

URL
Figure 1 (from left to right, top to bottom)
https://i.pinimg.com/736x/0d/5f/6f/0d5f6fc94deb23b2f19a1191683d8a9a.jpg
https://i.pinimg.com/736x/44/fa/77/44fa772708699f15804227777efb77c9.jpg
https://i.pinimg.com/736x/06/44/f2/0644f261057e108731cf769a180e1c75.jpg
https://i.pinimg.com/736x/20/c7/c8/20c7c83caaee3b743ce5f48e4757d515.jpg
https://i.pinimg.com/736x/7c/94/17/7c94174a952b1c630bb046b8ff833656.jpg
https://i.pinimg.com/1200x/63/75/dd/6375ddbf54ded4bdfe94c5ee62a44674.jpg
https://i.pinimg.com/736x/77/c1/cd/77c1cd929d545575e4cc9389d772638e.jpg
Figure 2 (from left to right, top to bottom)
https://i.pinimg.com/736x/00/21/3b/00213b38417aece9c3a7c2b1f20f279b.jpg
https://i.pinimg.com/1200x/11/c4/77/11c477e37102ced14faefdf5416c5174.jpg
Figure 3 (from left to right, top to bottom)
https://i.pinimg.com/1200x/20/9d/28/209d2847f56459d69984d10bec8880c6.jpg
https://i.pinimg.com/1200x/b3/93/71/b39371e2f6c19faa91ab5048dd6b4a6f.jpg
https://i.pinimg.com/736x/81/34/c2/8134c293c85b9f8a986c301a3a16734f.jpg
https://i.pinimg.com/1200x/60/b6/ce/60b6ce50b320ad83ddc716f59f04acee.jpg
Figure 4 (from left to right, top to bottom)
https://i.pinimg.com/originals/f4/0e/4f/f40e4f880d632c1c397a25c07b60d9bf.jpg
https://i.pinimg.com/originals/31/ce/21/31ce2169001832739 6194e5fae0e7186.jpg
https://i.pinimg.com/originals/c0/9d/3e/c09d3e1e67d47cc8d51e89ed4f6fbab5.jpg
https://i.pinimg.com/originals/98/ed/50/98ed50a28b0dcf44138d481375947a48.jpg
https://i.pinimg.com/originals/d0/fe/e6/d0fee6dc4b126beb9350e8875292d5f6.webp
https://i.pinimg.com/originals/89/ee/45/89ee45fa3ec303d6697f1162bfd84b66.jpg
https://i.pinimg.com/originals/de/92/47/de924728b927fdf5e79093f82c0a39ee.jpg
https://i.pinimg.com/originals/40/eb/47/40eb47fffb619192229f6063da70a895.jpg
https://i.pinimg.com/originals/72/f5/3a/72f53ae9719e425a9ec991ffda952df7.jpg
https://i.pinimg.com/originals/e9/8c/77/e98c771bc77f1c4204c228f58e79fd19.jpg
https://i.pinimg.com/originals/63/8f/77/638f776120f1d8853fd7ba6598be423d.jpg
https://i.pinimg.com/originals/af/fa/d4/affad41fe147de3f687e0b35309d33a7.jpg
https://i.pinimg.com/originals/15/87/0e/15870e4ad351967850359021 1434b296.jpg
https://i.pinimg.com/originals/d3/fa/77/d3fa77b346124c1fabfbaf7b73ee9a07.jpg
https://i.pinimg.com/originals/c5/63/a5/c563a5d9dd5681195af822a54d0fa9d6.jpg
https://i.pinimg.com/originals/e9/67/25/e96725f368f82c6982fc241a59439a95.jpg
https://i.pinimg.com/originals/7c/d4/89/7cd489ad783fcb1838f802b977df02ef.jpg
https://i.pinimg.com/originals/63/8f/77/638f776120f1d8853fd7ba6598be423d.jpg
https://i.pinimg.com/originals/e1/08/59/e10859fd26c1a285dd3cfba17c074e35.jpg
https://i.pinimg.com/originals/8a/52/af/8a52af98253decf83c0c2dfd4b1da0e3.jpg
https://i.pinimg.com/originals/dc/5e/81/dc5e81094a160103f05a4376159241c8.jpg
https://i.pinimg.com/originals/dc/5e/81/dc5e81094a160103f05a4376159241c8.jpg
https://i.pinimg.com/originals/ee/69/a9/ee69a9ecd73c2a572836e87ca4ca9c70.jpg
https://i.pinimg.com/originals/de/c3/76/dec3765bffeb8d9f761daf472270eab9.jpg
Figure 5 (from left to right, top to bottom)
https://i.pinimg.com/originals/2b/78/ed/2b78ed66f91a7f785c649e9b1907b53e.jpg
https://i.pinimg.com/originals/19/f2/fd/19f2fdd0ac564a26ebe534cf4a27814b.jpg
https://i.pinimg.com/originals/13/a2/c2/13a2c2c636b305e16445fb66cf936a5c.png
https://i.pinimg.com/originals/ad/4f/4d/ad4f4dd8f47301c73ad95a96b970fca6.jpg
https://i.pinimg.com/originals/29/37/10/29371102aef378ab6041db6ed18eded4f.jpg
https://i.pinimg.com/originals/55/e2/56/55e25640577f9786303e568dd93ec869.jpg

Table 10. A list of URLs of the images used in Fig. 1 to Fig. 5.

URL
Figure 11 (from left to right, top to bottom)
https://i.pinimg.com/originals/3f/f8/77/3ff8771991fba2c801c689ec348610b8.jpg
https://i.pinimg.com/originals/9b/d0/9f/9bd09f87f4c99947797da187f404bae0.jpg
https://i.pinimg.com/originals/9b/6c/23/9b6c231621d623040fce13d4d9a78163.jpg
https://i.pinimg.com/originals/47/98/64/47986441ac666239db471100a58ebd37.jpg
https://i.pinimg.com/originals/45/41/d1/4541d102732a19578dce9ae98049b0ed.jpg
https://i.pinimg.com/originals/5d/6f/c6/5d6fc6b827eb884d885d2532f0c1b9a4.jpg
https://i.pinimg.com/originals/81/cc/12/81cc1200a18f72a1c8860c35096dff76.jpg
https://i.pinimg.com/originals/0f/95/53/0f95531d3b18fd360bcbc9b7fc545540.jpg
https://i.pinimg.com/originals/31/ce/21/31ce21690018327396194e5fae0e7186.jpg
https://i.pinimg.com/originals/0f/14/69/0f146906dee51c92ac29804e368c589c.jpg
https://i.pinimg.com/originals/72/a2/1d/72a21d4c86b3df20523fff06910c5f44.jpg
https://i.pinimg.com/originals/75/1a/4b/751a4b19a1ff557bcf765bc44ea91dbe.png
https://i.pinimg.com/originals/a4/0d/2f/a40d2f99045779492853c03dd88903b4.jpg
https://i.pinimg.com/originals/d7/46/87/d74687af5a0e39b603d55d09d127a7ca.jpg
https://i.pinimg.com/originals/ba/7d/8d/ba7d8dc8d6895f2bdf503c04de570dba.png
https://i.pinimg.com/originals/5d/b0/0a/5db00afb28c4c72ad2f649fa7efa85ca.jpg
https://i.pinimg.com/originals/f3/57/ba/f357ba77afc0714286d81f0f580b19b0.jpg
https://i.pinimg.com/originals/12/2d/ef/122def22e2a6ed52cd333a1850c5886e.jpg
https://i.pinimg.com/originals/7c/94/17/7c94174a952b1c630bb046b8ff833656.jpg
https://i.pinimg.com/originals/ff/97/b4/ff97b457014484efacf1bb2ca078ca06.jpg
https://i.pinimg.com/originals/fc/b8/b3/fcb8b3dc9078383e817080cbb09c9890.jpg
https://i.pinimg.com/originals/78/74/ef/7874ef16f80d26826438090d091a2d0c.jpg
https://i.pinimg.com/originals/25/ec/e8/25ece8433cfddea6be0c2ca450c91f21.jpg
https://i.pinimg.com/originals/c5/3e/31/c53e319516e386a9dec92bc4105722c0.jpg
https://i.pinimg.com/originals/cd/6a/3b/cd6a3b6db72be958eea7c74a74d305ac.jpg
Figure 12 (from left to right, top to bottom)
https://i.pinimg.com/originals/6f/b5/34/6fb534c5e5d9870e534e9e90bb3be28e.png
https://i.pinimg.com/originals/6a/ec/77/6aec7735462214b968172d5351938598.jpg
https://i.pinimg.com/originals/76/12/d5/7612d5b8860edad3189a1c4799e07071.jpg
https://i.pinimg.com/originals/47/98/64/47986441ac666239db471100a58ebd37.jpg
https://i.pinimg.com/originals/45/41/d1/4541d102732a19578dce9ae98049b0ed.jpg
https://i.pinimg.com/originals/62/a5/a4/62a5a4130adb5df7d83144283623876a.jpg
https://i.pinimg.com/originals/99/9a/93/999a93a2e44612033432acadac532985.jpg
https://i.pinimg.com/originals/18/21/2e/18212ec316a3f47330687ed5afd754c0.jpg
https://i.pinimg.com/originals/47/98/64/47986441ac666239db471100a58ebd37.jpg
https://i.pinimg.com/originals/45/41/d1/4541d102732a19578dce9ae98049b0ed.jpg
https://i.pinimg.com/originals/c6/2b/fe/c62bfe2a2d38c76104d93d503366f47c.jpg
https://i.pinimg.com/originals/48/bf/68/48bf685481c952a38cba8347e06c67e5.jpg
https://i.pinimg.com/originals/dc/e7/57/dce75798bee0c71c8baa81ab6a31eaa3.jpg
https://i.pinimg.com/originals/47/98/64/47986441ac666239db471100a58ebd37.jpg
https://i.pinimg.com/originals/45/41/d1/4541d102732a19578dce9ae98049b0ed.jpg
https://i.pinimg.com/originals/5f/61/54/5f6154942ef97c36a435085139146bd8.jpg
https://i.pinimg.com/originals/52/77/f0/5277f03cbc497552d37de2c8ca69ac16.jpg
https://i.pinimg.com/originals/5a/e2/35/5ae2356dc37e4c1a50abe5c3c55e11f8.jpg
https://i.pinimg.com/originals/47/98/64/47986441ac666239db471100a58ebd37.jpg
https://i.pinimg.com/originals/45/41/d1/4541d102732a19578dce9ae98049b0ed.jpg
Figure 13 (from left to right, top to bottom)
https://i.pinimg.com/originals/c9/51/c6/c951c608b1e8dd5c6ad68e30eb2be735.jpg
https://i.pinimg.com/originals/0f/14/69/0f146906dee51c92ac29804e368c589c.jpg
https://i.pinimg.com/originals/52/77/f0/5277f03cbc497552d37de2c8ca69ac16.jpg
https://i.pinimg.com/originals/dc/e7/57/dce75798bee0c71c8baa81ab6a31eaa3.jpg
https://i.pinimg.com/originals/78/74/ef/7874ef16f80d26826438090d091a2d0c.jpg
https://i.pinimg.com/originals/47/59/bf/4759bfbfa7ba31679973e5b056206493.jpg
https://i.pinimg.com/originals/e6/97/e7/e697e77765fd02f5190f19770ab007ef.jpg
https://i.pinimg.com/originals/d2/4f/20/d24f20fe2f2c8f9fd313358c2ab0ddc5.jpg
https://i.pinimg.com/originals/72/a2/1d/72a21d4c86b3df20523fff06910c5f44.jpg
https://i.pinimg.com/originals/75/1a/4b/751a4b19a1ff557bcf765bc44ea91dbe.png
https://i.pinimg.com/originals/0a/45/4e/0a454e7d5dce658ddac2c7baa2978cb9.jpg
https://i.pinimg.com/originals/d3/68/8b/d3688bc7945d997cedb6fe6289ddcaa4.jpg
https://i.pinimg.com/originals/4d/76/44/4d7644fea93ca1f5612d401ad405ea54.jpg
https://i.pinimg.com/originals/72/a2/1d/72a21d4c86b3df20523fff06910c5f44.jpg
https://i.pinimg.com/originals/75/1a/4b/751a4b19a1ff557bcf765bc44ea91dbe.png
https://i.pinimg.com/originals/62/88/be/6288be6450248c9ea58bb1220b27d2a4.jpg
https://i.pinimg.com/originals/e2/e4/0f/e2e40fa9847cb77fa2be0ff629eb8c59.jpg
https://i.pinimg.com/originals/06/f9/63/06f963624007af619d2e70b17d8397f9.jpg
https://i.pinimg.com/originals/72/a2/1d/72a21d4c86b3df20523fff06910c5f44.jpg
https://i.pinimg.com/originals/75/1a/4b/751a4b19a1ff557bcf765bc44ea91dbe.png

Table 11. A list of URLs of the images used in Fig. 11 to Fig. 13.