UFineBench: Towards Text-based Person Retrieval with Ultra-fine Granularity

Supplementary Material

Contents

In this supplementary material, we will 1) show the details of our proposed cross-modal identity classifying loss $\mathcal{L}_{cid}$, 2) show more results of the ablation study, and 3) show more specific and compared examples of our proposed UFineBench and existing other datasets [1, 2, 4].

1. Cross-Modal Identity Classifying

In the training phase of CFAM, we also propose the cross-modal identity classifying loss $\mathcal{L}_{cid}$ as a supplement to explicitly ensure that the representations of the same image/text pair are closely clustered together.

First, referring to [3], we revisit the traditional identity loss commonly used in the person re-identification task. Given the extracted embeddings $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^N$ and the identity labels $\mathcal{Y} = \{y_i\}_{i=1}^N$, the traditional identity loss can be computed by:

$$\mathcal{L}_{id} = \frac{1}{N}\sum_i -\log\left(\frac{\exp\left(\boldsymbol{W}_{y_i}^\top \boldsymbol{x}_i + b_{y_i}\right)}{\sum_j \exp\left(\boldsymbol{W}_j^\top \boldsymbol{x}_i + b_j\right)}\right), \quad (1)$$

where $\boldsymbol{W}_{y_i}$ and $\boldsymbol{W}_j$ denote the y_i-th and j-th column of classification weight matrix $\boldsymbol{W}$, y_i indicates the identity label of $\boldsymbol{x}_i$, and b_{y_i} and b_j represent the y_i-th and j-th element of bias vector.

However, this loss only performs identity clustering on individual modalities, lacking interaction across different modalities and unable to conduct feature clustering in a shared cross-modal space. Therefore, we propose the cross-modal identity classifying loss, which not only considers the cross-modal correlation but also deeply mines hard negative unmatched sample pairs.

Given the extracted visual embeddings $\mathcal{V} = \{\boldsymbol{v}_i\}_{i=1}^N$ and textual embeddings $\mathcal{T} = \{\boldsymbol{t}_i\}_{i=1}^N$, for each $\boldsymbol{v}_i$ within $\mathcal{V}$, we sample the unpaired textual embedding which owns the highest similarity with this $\boldsymbol{v}_i$ as the negative. Also, we sample one hard negative visual embedding for each $\boldsymbol{t}_i$ within $\mathcal{T}$ in the same way. Specially, we add an extra identity label as the unmatched label, that is, if the original identity labels are from M classes, the $(M+1)$-th identity will be set as the unmatched label. Through this approach, we obtain $|N|$ original positive pairs with original identity labels and $2|N|$ negative pairs with an unmatched label, denoted as $|B|$ pairs with the identity labels $\{y_i\}_{i=1}^B$.

Then, for the $|B|$ pairs $\{\boldsymbol{v}_i, \boldsymbol{t}_i, y_i\}_{i=1}^B$, we first concatenate the visual embedding $\boldsymbol{v}_i$ and textual embedding $\boldsymbol{t}_i$ to form the cross-modal embedding $\boldsymbol{z}_i$. Then, for $\{\boldsymbol{z}_i\}_{i=1}^B$ with the identity labels $\{y_i\}_{i=1}^B$, the cross-modal identity classifying loss can be computed by:

$$\mathcal{L}_{cid} = \frac{1}{N}\sum_i -\log\left(\frac{\exp\left(\boldsymbol{W}_{y_i}^\top mlp(\boldsymbol{z}_i) + b_{y_i}\right)}{\sum_j \exp\left(\boldsymbol{W}_j^\top mlp(\boldsymbol{z}_i) + b_j\right)}\right), \quad (2)$$

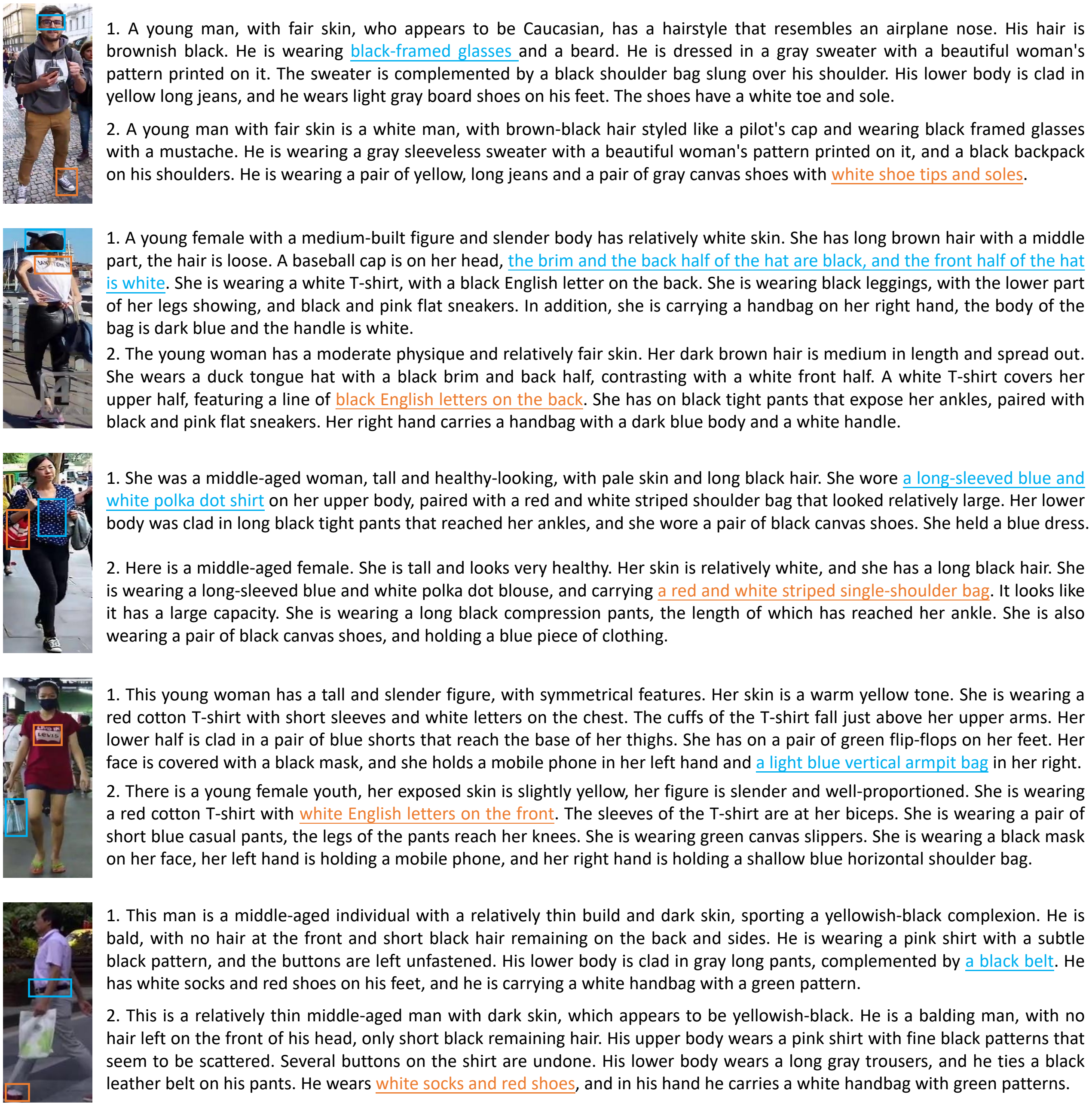
where mlp denotes an MLP layer consisted of a Linear layer, a LayerNorm, a GELU activation to fuse the cross-modal embeddings more deeply.

No.	Components			CUHK-PEDES				
	share	depth	queries	R@1	R@5	R@10	mAP	mSD
0		1	16	71.17	87.83	92.72	63.83	49.00
1		2	16	71.05	87.56	92.58	63.79	49.10
2		3	16	70.96	87.69	92.31	63.56	48.79
3		4	16	71.72	87.76	92.75	64.13	49.13
4	✓	**2**	**4**	**72.87**	**88.61**	92.87	**64.92**	**50.20**
5	✓	2	8	71.72	88.34	93.00	64.32	49.64
6	✓	2	12	71.61	88.30	92.72	64.27	49.72
7	✓	2	16	71.83	88.43	93.15	64.39	49.52
8	✓	2	20	72.08	88.48	93.02	64.50	49.51
9	✓	3	4	72.09	88.48	93.05	64.44	49.68
10	✓	3	8	71.59	88.61	**93.34**	64.06	49.06
11	✓	3	12	72.34	88.50	93.00	64.40	49.38
12	✓	3	16	71.85	88.32	92.95	64.23	49.40
13	✓	3	20	72.24	88.55	82.92	64.40	49.64
14	✓	4	4	71.41	88.30	93.02	64.17	49.73
15	✓	4	8	72.13	88.56	93.02	64.32	49.35
16	✓	4	12	71.61	88.32	92.98	64.37	49.63
17	✓	4	16	72.04	88.58	92.97	64.44	49.66
18	✓	4	20	72.06	88.42	93.00	64.41	49.80

Table 1. Ablation study on some components of CFAM. The "share" denotes whether the granularity decoder is shared across modalities. The "depth" denotes the number of transformer blocks in the granularity decoder. The "queries" represents the number of query tokens used to extract fine-grained information.

2. More Results of Ablation Study

We conduct an ablation experiment to study the influence of whether the granularity decoder is shared across modalities, the number of transformer blocks in the granularity decoder and the number of query tokens. The models are trained and evaluated on the CUHK-PEDES dataset [2]. According to the results in the Table 1, we can draw two conclusions as follows. First, compared to an unshared granularity decoder (No.1, No.2, No.3), a shared granularity decoder (No.7, No12, No.17) can bring about a significant improvement in performance. Second, when the number of transformer blocks in the granularity decoder and query tokens are 2 and 16 (No.4), respectively, the best performance is obtained, achieving 72.87% rank-1, 64.92% mAP, and 50.20% mSD, which is set as the default in CFAM.



1. A young man, with fair skin, who appears to be Caucasian, has a hairstyle that resembles an airplane nose. His hair is brownish black. He is wearing black-framed glasses and a beard. He is dressed in a gray sweater with a beautiful woman's pattern printed on it. The sweater is complemented by a black shoulder bag slung over his shoulder. His lower body is clad in yellow long jeans, and he wears light gray board shoes on his feet. The shoes have a white toe and sole.

2. A young man with fair skin is a white man, with brown-black hair styled like a pilot's cap and wearing black framed glasses with a mustache. He is wearing a gray sleeveless sweater with a beautiful woman's pattern printed on it, and a black backpack on his shoulders. He is wearing a pair of yellow, long jeans and a pair of gray canvas shoes with white shoe tips and soles.

1. A young female with a medium-built figure and slender body has relatively white skin. She has long brown hair with a middle part, the hair is loose. A baseball cap is on her head, the brim and the back half of the hat are black, and the front half of the hat is white. She is wearing a white T-shirt, with a black English letter on the back. She is wearing black leggings, with the lower part of her legs showing, and black and pink flat sneakers. In addition, she is carrying a handbag on her right hand, the body of the bag is dark blue and the handle is white.

2. The young woman has a moderate physique and relatively fair skin. Her dark brown hair is medium in length and spread out. She wears a duck tongue hat with a black brim and back half, contrasting with a white front half. A white T-shirt covers her upper half, featuring a line of black English letters on the back. She has on black tight pants that expose her ankles, paired with black and pink flat sneakers. Her right hand carries a handbag with a dark blue body and a white handle.

1. She was a middle-aged woman, tall and healthy-looking, with pale skin and long black hair. She wore a long-sleeved blue and white polka dot shirt on her upper body, paired with a red and white striped shoulder bag that looked relatively large. Her lower body was clad in long black tight pants that reached her ankles, and she wore a pair of black canvas shoes. She held a blue dress.

2. Here is a middle-aged female. She is tall and looks very healthy. Her skin is relatively white, and she has a long black hair. She is wearing a long-sleeved blue and white polka dot blouse, and carrying a red and white striped single-shoulder bag. It looks like it has a large capacity. She is wearing a long black compression pants, the length of which has reached her ankle. She is also wearing a pair of black canvas shoes, and holding a blue piece of clothing.

1. This young woman has a tall and slender figure, with symmetrical features. Her skin is a warm yellow tone. She is wearing a red cotton T-shirt with short sleeves and white letters on the chest. The cuffs of the T-shirt fall just above her upper arms. Her lower half is clad in a pair of blue shorts that reach the base of her thighs. She has on a pair of green flip-flops on her feet. Her face is covered with a black mask, and she holds a mobile phone in her left hand and a light blue vertical armpit bag in her right.

2. There is a young female youth, her exposed skin is slightly yellow, her figure is slender and well-proportioned. She is wearing a red cotton T-shirt with white English letters on the front. The sleeves of the T-shirt are at her biceps. She is wearing a pair of short blue casual pants, the legs of the pants reach her knees. She is wearing green canvas slippers. She is wearing a black mask on her face, her left hand is holding a mobile phone, and her right hand is holding a shallow blue horizontal shoulder bag.

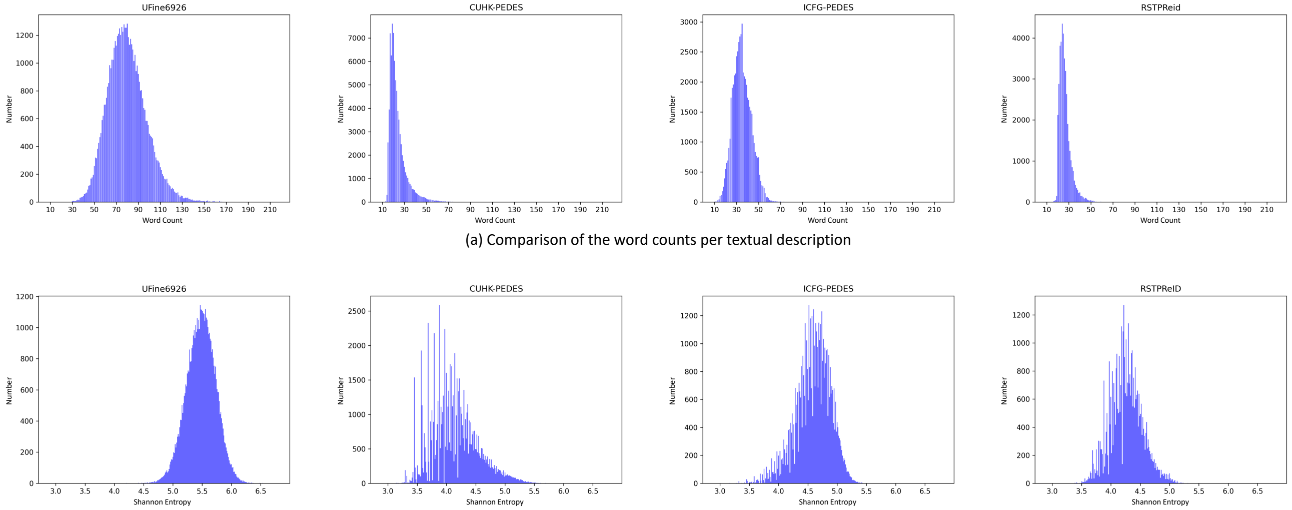
1. This man is a middle-aged individual with a relatively thin build and dark skin, sporting a yellowish-black complexion. He is bald, with no hair at the front and short black hair remaining on the back and sides. He is wearing a pink shirt with a subtle black pattern, and the buttons are left unfastened. His lower body is clad in gray long pants, complemented by a black belt. He has white socks and red shoes on his feet, and he is carrying a white handbag with a green pattern.

2. This is a relatively thin middle-aged man with dark skin, which appears to be yellowish-black. He is a balding man, with no hair left on the front of his head, only short black remaining hair. His upper body wears a pink shirt with fine black patterns that seem to be scattered. Several buttons on the shirt are undone. His lower body wears a long gray trousers, and he ties a black leather belt on his pants. He wears white socks and red shoes, and in his hand he carries a white handbag with green patterns.

Figure 1. Some examples of our proposed UFine6926. Every image has two different fine-grained textual descriptions that describes the person's apperance detailedly. Some fine-grained features are highlighted in blue or orange boxes and texts accordingly.

3. More Examples of UFineBench

Ultra Fine-grained UFine6926. We have shown more representative examples of our proposed ultra fine-grained UFine6926 in Figure 1. As we can see, each person image is annotated with two different textual descriptions that describes the person's appearance detailedly. Even if the external characteristics of certain persons in the images are very subtle, our textual descriptions do not ignore them like the previous datasets [1, 2, 4] and portrays these fine-grained features accurately. These fine-grained features are highlighted in blue or orange boxes and texts in the Figure 1. For instance, in the bottom example, the area of the person's shoes is very inconspicuous, yet our textual description accurately identifies this area and describes it as "white socks and red shoes." Meanwhile, we conduct a statistical comparison of the word counts per textual description in our UFine6926 with those in other datasets, as illustrated in Figure 2 (a). By examining the distribution, we can



(a) Comparison of the word counts per textual description

(b) Comparison of the Shannon entropy per textual description

Figure 2. Statistical comparison of the word counts per textual description in our UFine6926 with those in other datasets [1, 2, 4].

Benchmark	MaxEnt	MinEnt	AverEnt	AllEnt
CUHK-PEDES	5.901	3.057	4.140	6.956
ICFG-PEDES	5.604	3.122	4.593	6.412
RSTPReid	5.496	2.914	4.265	6.448
UFine6926	**6.621**	**4.201**	**5.480**	**7.465**

Table 2. Comparison of entropy-based metrics for each dataset. "MaxEnt", "MinEnt", "AverEnt" represent text-level maximum, minimum and average entropy, respectively. "AllEnt" is the dataset-level entropy calculated by aggregating all the texts in the dataset. The higher the entropy, the richer the information.

observe that the word count for UFine6926 is generally centered around 80, while simultaneously, CUHK-PEDES [2] and RSTPReid [4] are both roughly centered around 25, and ICFG-PEDES [1] is centered around 30. It is evident that the level of textual detail in our UFine6926 is higher than that in all other datasets by a significant margin. At the same time, in information theory, Shannon entropy is often used to measure the amount of information in a system. For a text, we can treat each word as an event, calculate the probability of each word occurring, and finally calculate the entropy of the entire text based on Shannon's entropy formula. We conduct a statistical comparison of the Shannon entropy per textual description in our UFine6926 with those in other datasets, as illustrated in Figure 2 (b). Meanwhile, we also calculate the specific Shannon entropy metrics for each dataset, as shown in Table 2. The higher the entropy, the greater the amount of information in the text. From the distribution and metrics, we can see that our UFine6926 has significantly richer textual information than other datasets.

Three Cross Settings in UFine3C. Our proposed UFine3C evaluation set has cross domains, cross text granularity and cross text styles, which is more representative of the challenges faced in real scenarios. (1) *Our UFine3C spans across various domains.* As shown in Figure 3, the images in UFine3C have significant variations in resolution, illumination and shooting scenes, which is very close to the situation of images obtained in real scenarios. (2) *Our UFine3C spans across various text granularity.* As shown in Figure 4, on the left, the word counts distribution of UFine3C is spanning from coarse-grained to fine-grained. Meanwhile, on the right, according to the text granularity, the texts can be categorized into the fine-grained, the medium-grained and the coarse-grained, respectively. This represents the inconsistency of granularity within the query texts in real scenarios. (3) *Our UFine3C spans across various text styles.* As shown in Figure 5, each image has multiple query texts with different styles, simulating the language expression styles of different individuals in practice.

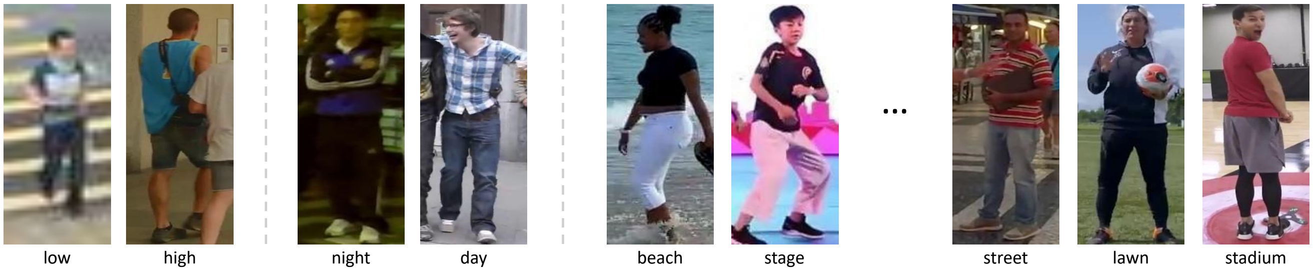
References

[1] Zefeng Ding, Changxing Ding, Zhiyin Shao, and Dacheng Tao. Semantically self-aligned network for text-to-image part-aware person re-identification. *arXiv preprint arXiv:2107.12666*, 2021. 1, 2, 3

[2] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *CVPR*, pages 1970–1979, 2017. 1, 2, 3

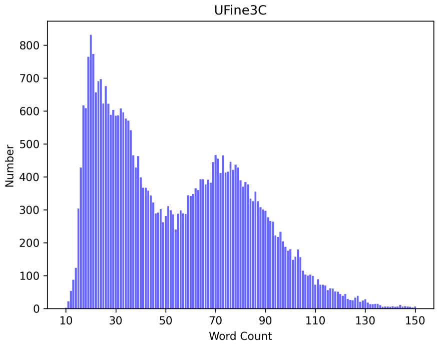
[3] Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *ECCV*, pages 686–701, 2018. 1

[4] Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua. Dssl: Deep surroundings-person separation learning for text-based person retrieval. In *ACM MM*, pages 209–217, 2021. 1, 2, 3



(a) Resolution (b) Illumination (c) Scene

Figure 3. Our UFine3C spans across various **domains** such as image resolution, illumination, and shooting scenes.



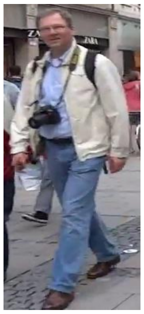
Fine-grained: This man is middle-aged and has a tall, slender physique. His arms and legs are long and lean, and his skin is very fair. He has black hair and is wearing a black top hat and sunglasses. He was dancing with exaggerated movements, spreading his hands and tiptoeing on one foot. He is wearing a white V-neck top underneath a black willow top, with white stitching on the right arm. His lower body is clad in long, black, slim-fitting pants that have reached the ankle position. He has a black rivet belt tied around his waist, and is wearing white socks and black leather shoes on his feet.

Medium-grained: The woman has her black hair tied back in a ponytail and is wearing a sleek black blouson jacket, which is paired with matching black trousers. She's also carrying a backpack slung over her shoulder, completing her stylish and practical outfit.

Coarse-grained: A woman wearing a dark black jacket with buttons, a pair of black and white pants and a pair of white shoes.

(a) Word Counts Distribution (b) Textual Descriptions with Different Granularity

Figure 4. Our UFine3C spans across various **text granularity** from coarse-grained to fine-grained. The word counts distribution is shown in (a). The (b) illustrates some specific examples representing three different text granularity conditions.



Style1: This man is middle-aged with fair skin and a tall, proportionate build. His short yellow hair is neatly styled. He's wearing a blue shirt underneath a beige long-sleeved jacket that reaches his buttocks. The jacket is the perfect length, and he's paired it with long blue jeans that fall below his ankles. On his feet, he wears brown leather shoes. A black backpack rests comfortably on his back, and a black DSLR camera hangs around his neck.

Style2: A middle-aged man with fair skin stands tall an proportional, sporting neat yellow hair and a stylish blue shirt under a beige long-sleeved jacket that falls to his buttocks. Paired with long blue jeans that trail below his ankles, he completes his look with brown leather shoes, a black backpack, and a black DSLR camera hanging around his neck.

Style3: This is a middle-aged man with fair skin, average height, and well-proportioned build. He has short yellow hair. His upper body is wearing a blue shirt, with a beige long-sleeved jacket on top, which reaches just below his butt. His lower body is wearing a long blue jeans that goes down to just above his ankles. The man's feet are wearing a pair of brown leather shoes. He is carrying a black backpack on his back. A black single-lens reflex camera hangs around his chest.

Style4: This man is middle-aged, with a fair complexion and a well-proportioned build. He has short yellow hair and is wearing a blue shirt underneath a beige long-sleeved jacket that reaches just below his butt. His lower body is clad in long blue jeans that stop just above his ankles. On his feet, he wears brown leather shoes. He carries a black backpack on his back and has a black single-lens reflex camera hanging around his chest.

Style5: A middle-aged man with fair skin stands before us. His height is average, but his build is well-proportioned. His short yellow hair falls neatly across his forehead. His outfit consists of a blue shir worn on his upper body, complemented by a beige long-sleeved jacket that stops just below his butt. His lower half is covered by long blue jeans that extend down to just above his ankles. Brown leather shoes grace his feet, providing a touch of sophistication to his overall look. Carrying a black backpack on his back, he also sports a black single-lens reflex camera hanging around his chest, indicating a passion for photography.

Style6: The man is tall and slender, with a build that's well-proportioned. He has fair skin and short, yellow hair that's styled neatly. He's wearing a blue shirt underneath a beige long-sleeved jacket that fits him perfectly and reaches his buttocks. The jacket is paired with long blue jeans that fall below his ankles, and he's wearing brown leather shoes. A black backpack rests comfortably on his back, and a black DSLR camera hangs around his neck.

Figure 5. Our UFine3C spans across various **text styles**, simulating the language expression styles of different individuals.