Contents

A. Eevee dataset

A.1 Raw data

As shown in Figure 1, Figure 2 and Figure 3, we present raw data samples for the dresses, lower body, and upper body categories of the Eevee dataset, respectively. A sample's raw data consists of a full-shot person video , a close-up person video, a garment image, a garment detail image, and a person image.

A.2 Parsing annotation

To enhance the usability of the dataset for various downstream tasks, we provide a comprehensive suite of annotations:

1. We utilize the multimodal large language model, Qwen-VL-Max, to generate detailed textual descriptions of the garments and to classify them into their respective categories.
2. For full-shot virtual try-on videos and person images, we use OpenPose to obtain human parsing results and then generate the mask. For close-up virtual try-on videos, where obtaining human parsing results is often difficult, we instead use Grounded SAM-2 to perform semantic segmentation on the garment region and use that as the mask.
3. To adapt to other virtual try-on models, we use AniLines to extract garment contour maps and Detectron2 to obtain DensePose UV coordinates for the human body.

As shown in Figure 4, Figure 5, and Figure 6, we present complete data samples from the Eevee dataset for the dresses, lower-body, and upper-body categories, respectively. Each complete sample includes a full-shot person video along with its corresponding masked video and DensePose UV coordinates, a close-up person video along with its corresponding masked video and DensePose UV coordinates, a garment image along with its corresponding garment detail textual description and line map, a person image along with its corresponding masked image, and a garment detail image.

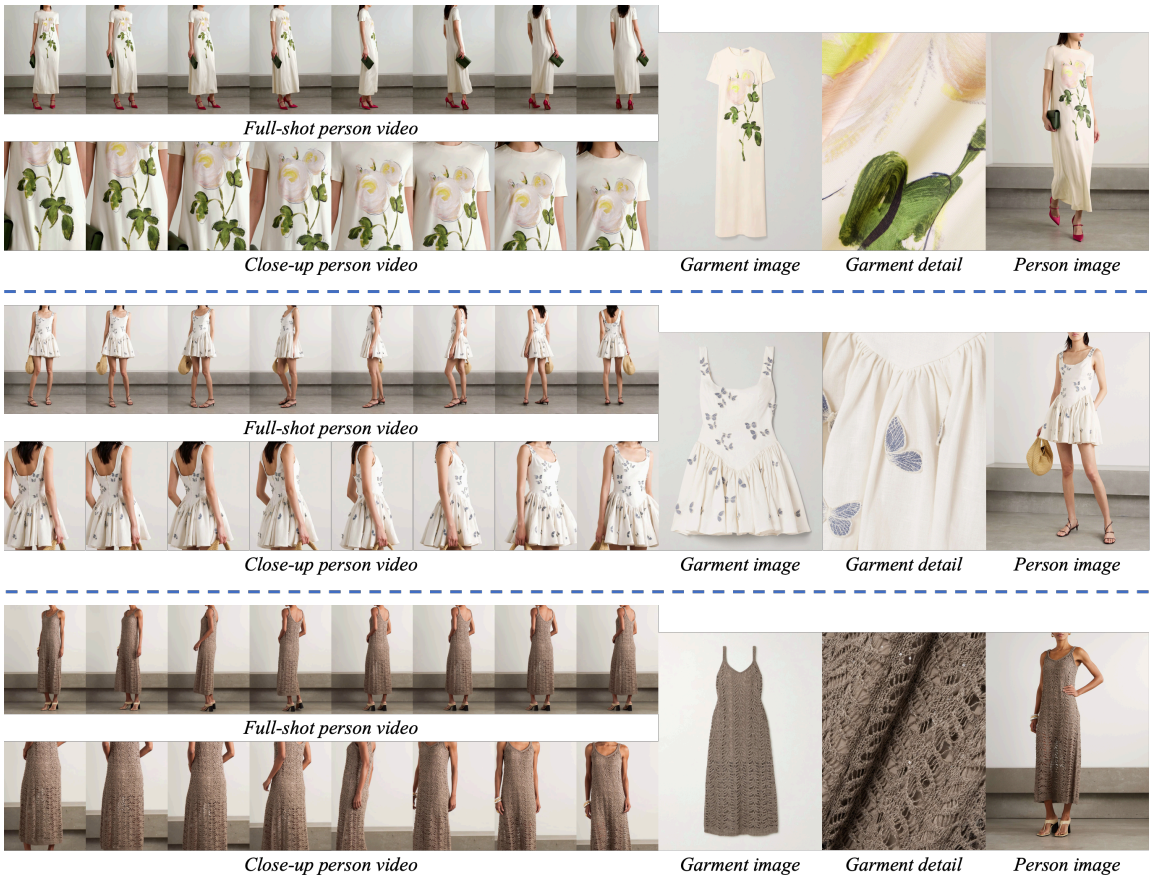

Figure 1: Raw data samples from the dresses category.

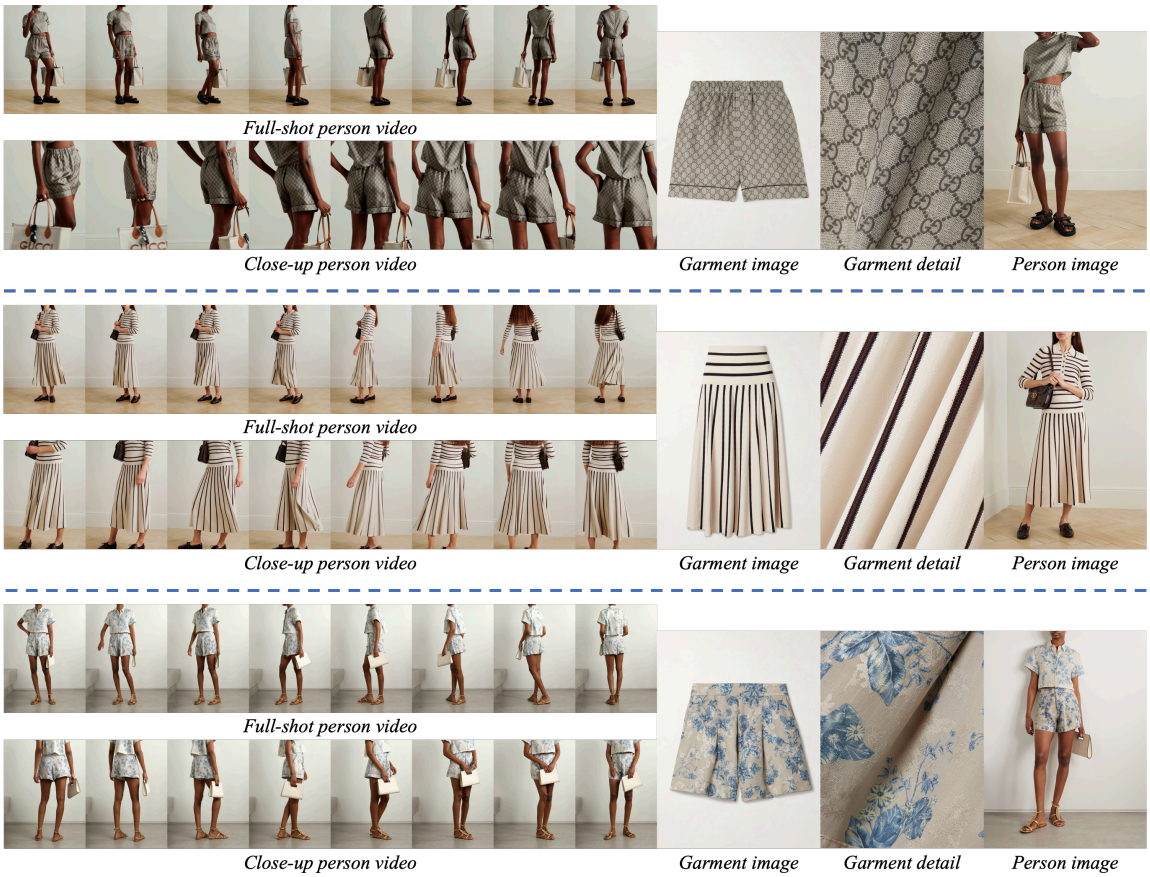

Figure 2: Raw data samples from the lower body category.

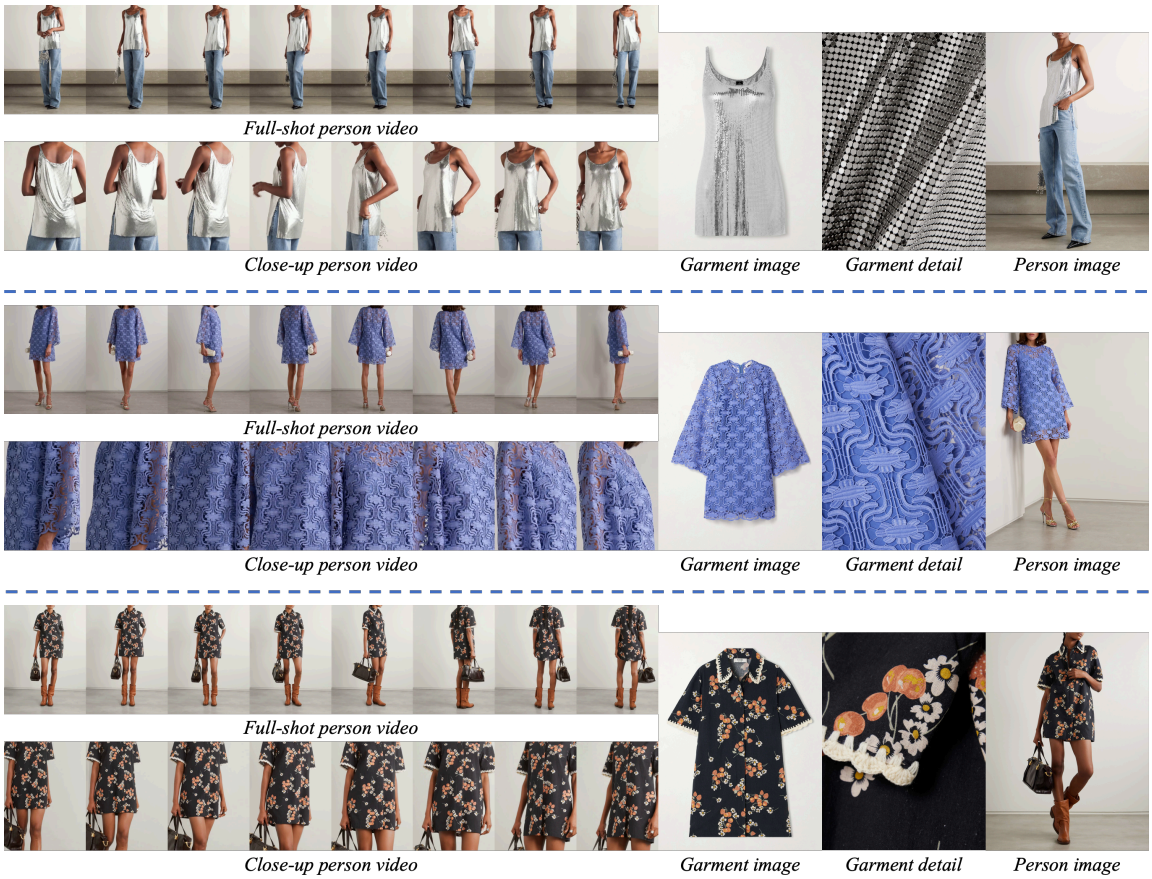

Figure 3: Raw data samples from the upper body category.

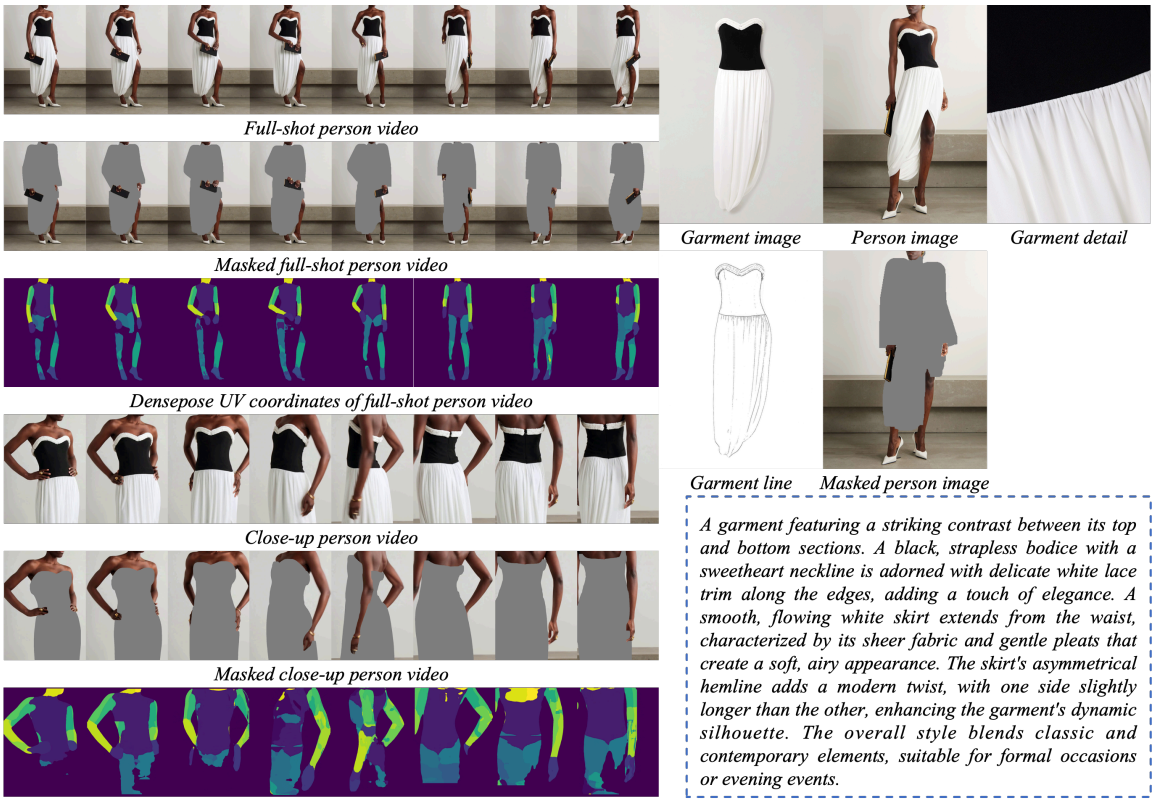

Figure 4: Complete data samples from the dresses category.

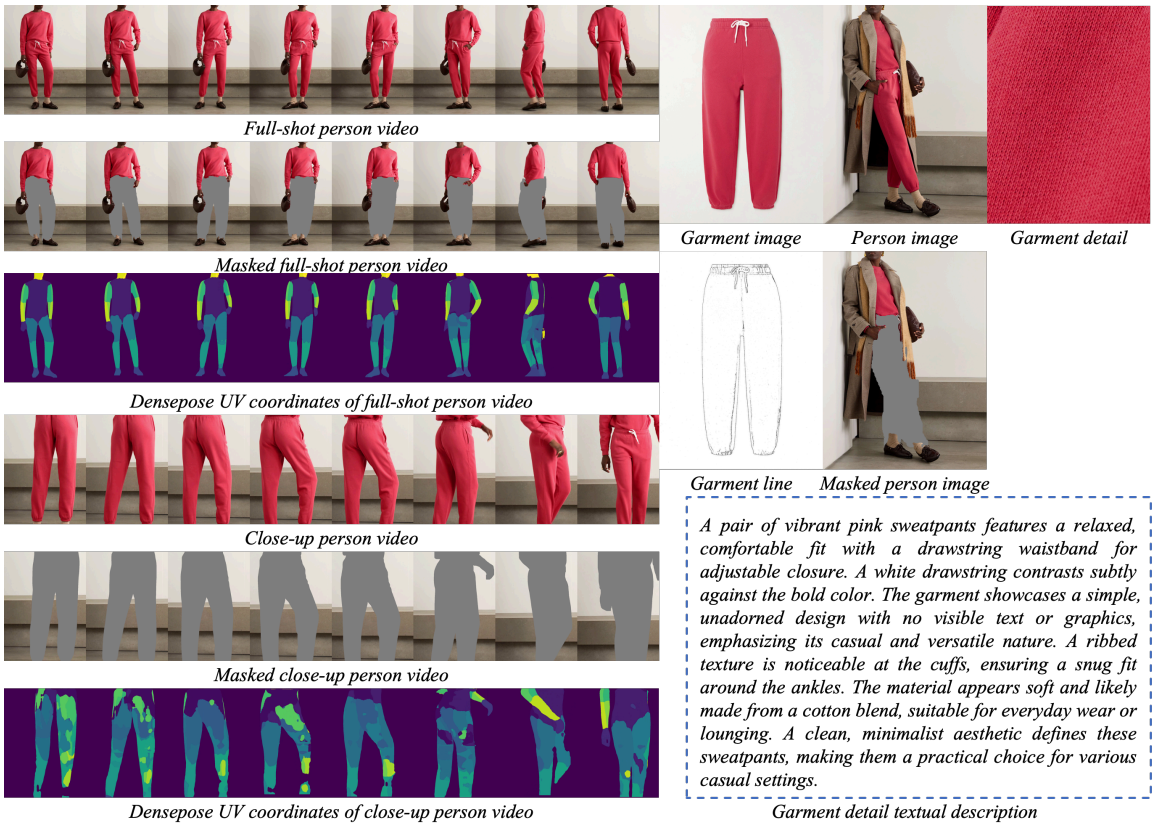

Figure 5: Complete data samples from the lower body category.

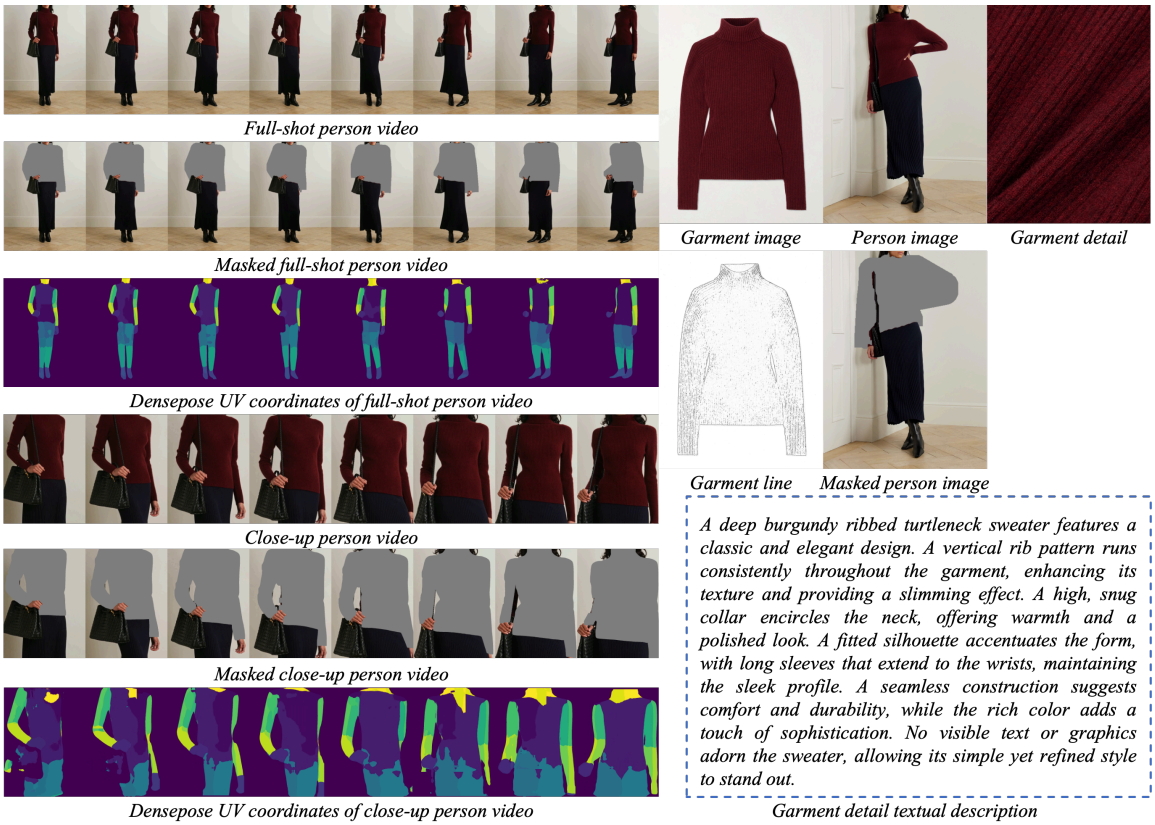

Figure 6: Complete data samples from the upper body category.

B. Additional quantitative results

The Eevee dataset is categorized into three distinct subsets: dresses, lower body, and upper body. We report comprehensive quantitative metrics for each category individually to ensure a granular analysis of model performance.

Our evaluation framework includes VFID (ResNeXt) and VFID (I3D) to assess video realism, VGID for garment consistency, and standard image quality metrics such as LPIPS and SSIM for paired structural similarity. Furthermore, we employ VBench to holistically evaluate the generated videos across four dimensions: Subject Consistency, Background Consistency, Aesthetic Quality, and Imaging Quality.

B.1 Dresses

Resolution	Type	Method	Unpaired			Paired				
			$VFID_R$↓	$VFID_I$↓	VGID↑	$VFID_R$↓	$VFID_I$↓	VGID↑	SSIM↑	LPIPS↓
1088 × 816	Full-shot	ViViD	1.015	19.128	0.514	0.792	15.584	0.521	0.835	0.145
		MagicTryon	**0.368**	**16.419**	0.512	0.263	**10.748**	0.524	**0.872**	0.110
		VACE	0.565	17.928	0.522	0.488	12.585	0.525	0.855	0.107
		Ours	0.402	17.033	**0.524**	**0.149**	11.604	**0.534**	0.869	**0.099**
832 × 624	Full-shot	ViViD	0.644	18.356	0.506	0.381	14.049	0.508	0.831	0.128
		MagicTryon	**0.323**	**16.519**	0.520	0.239	**10.641**	0.516	**0.870**	**0.091**
		VACE	0.676	18.232	0.524	0.488	12.760	**0.528**	0.840	0.114
		Ours	0.399	16.826	**0.528**	**0.149**	11.480	0.527	0.860	0.092
1088 × 816	Close-up	ViViD	0.925	16.068	0.533	0.838	13.297	0.509	0.779	0.183
		MagicTryon	0.842	**15.350**	0.538	0.834	**10.676**	0.519	**0.803**	0.162
		VACE	1.217	17.385	0.533	0.431	11.538	0.515	0.774	0.172
		Ours	**0.731**	15.710	**0.540**	**0.318**	10.838	**0.524**	0.798	**0.154**
832 × 624	Close-up	ViViD	0.757	16.009	0.533	0.565	12.594	0.509	0.767	0.175
		MagicTryon	**0.665**	**15.437**	0.534	0.606	**10.507**	0.513	**0.787**	0.150
		VACE	1.525	17.305	0.545	0.642	12.035	0.519	0.752	0.176
		Ours	0.727	15.477	**0.548**	**0.318**	10.732	**0.526**	0.780	**0.148**

Table 1: Quantitative comparison results for the dresses category.

Resolution	Type	Method	Unpaired				Paired			
			Subject Consistency	Background Consistency	Aesthetic Quality	Imaging Quality	Subject Consistency	Background Consistency	Aesthetic Quality	Imaging Quality
1088 × 816	Full-shot	ViViD	0.959	0.959	0.511	0.718	0.962	0.960	0.520	0.717
		MagicTryon	0.965	0.954	0.515	0.703	0.967	0.955	0.527	0.704
		VACE	0.962	0.966	0.532	0.722	0.964	0.967	0.543	0.721
		Ours	0.965	0.970	0.528	0.719	0.965	0.970	0.539	0.717
832 × 624	Full-shot	ViViD	0.961	0.962	0.498	0.715	0.964	0.964	0.506	0.714
		MagicTryon	0.966	0.971	0.511	0.711	0.968	0.971	0.522	0.711
		VACE	0.963	0.965	0.525	0.723	0.965	0.967	0.538	0.721
		Ours	0.964	0.969	0.525	0.708	0.966	0.967	0.541	0.704
1088 × 816	Close-up	ViViD	0.943	0.950	0.476	0.676	0.945	0.952	0.488	0.672
		MagicTryon	0.949	0.957	0.496	0.669	0.950	0.958	0.508	0.663
		VACE	0.947	0.953	0.520	0.696	0.947	0.954	0.524	0.689
		Ours	0.951	0.956	0.520	0.692	0.955	0.956	0.525	0.683
832 × 624	Close-up	ViViD	0.941	0.947	0.476	0.681	0.943	0.948	0.488	0.679
		MagicTryon	0.948	0.939	0.495	0.664	0.948	0.939	0.504	0.659
		VACE	0.945	0.953	0.523	0.692	0.946	0.954	0.528	0.689
		Ours	0.945	0.957	0.515	0.674	0.947	0.957	0.520	0.669

Table 2: VBench evaluation results for the dresses category.

B.2 Lower body

Resolution	Type	Method	Unpaired		Paired			
			$VFID_R$↓	$VFID_I$↓	$VFID_R$↓	$VFID_I$↓	SSIM↑	LPIPS↓
1088 × 816	Full-shot	ViViD	0.510	14.191	0.502	10.217	0.909	0.087
		MagicTryon	**0.127**	**9.736**	0.064	5.442	0.929	0.080
		VACE	0.316	12.184	0.093	6.128	0.929	0.063
		Ours	0.135	9.752	**0.041**	**5.525**	**0.939**	**0.059**
832 × 624	Full-shot	ViViD	0.333	13.285	0.287	8.634	0.903	0.083
		MagicTryon	**0.118**	9.994	**0.034**	5.453	0.932	0.062
		VACE	0.468	12.749	0.159	6.829	0.920	0.068
		Ours	0.133	**9.713**	0.040	**5.526**	**0.936**	**0.053**
1088 × 816	Close-up	ViViD	1.497	16.402	1.014	14.139	0.813	0.164
		MagicTryon	0.637	**13.695**	0.403	10.990	**0.835**	0.150
		VACE	0.679	15.220	0.452	11.252	0.803	0.155
		Ours	**0.445**	14.276	**0.229**	**10.131**	0.831	**0.133**
832 × 624	Close-up	ViViD	1.210	15.470	0.741	12.988	0.805	0.155
		MagicTryon	0.510	**13.623**	0.283	10.647	**0.818**	0.139
		VACE	0.765	15.221	0.569	11.717	0.780	0.161
		Ours	**0.438**	14.094	**0.227**	**10.031**	0.815	**0.126**

Table 3: Quantitative comparison results for the lower body category.

Resolution	Type	Method	Unpaired				Paired			
			Subject Consistency	Background Consistency	Aesthetic Quality	Imaging Quality	Subject Consistency	Background Consistency	Aesthetic Quality	Imaging Quality
1088 × 816	Full-shot	ViViD	0.955	0.960	0.491	0.716	0.957	0.961	0.507	0.715
		MagicTryon	0.960	0.952	0.491	0.695	0.962	0.954	0.507	0.697
		VACE	0.957	0.964	0.500	0.714	0.959	0.965	0.518	0.713
		Ours	0.961	0.967	0.501	0.715	0.960	0.970	0.516	0.715
832 × 624	Full-shot	ViViD	0.956	0.963	0.477	0.709	0.959	0.964	0.493	0.710
		MagicTryon	0.961	0.967	0.489	0.704	0.962	0.968	0.503	0.704
		VACE	0.958	0.964	0.495	0.711	0.960	0.966	0.514	0.710
		Ours	0.960	0.965	0.500	0.698	0.961	0.965	0.515	0.698
1088 × 816	Close-up	ViViD	0.933	0.950	0.459	0.669	0.936	0.952	0.463	0.669
		MagicTryon	0.941	0.954	0.475	0.657	0.943	0.955	0.481	0.659
		VACE	0.938	0.951	0.485	0.689	0.939	0.951	0.488	0.685
		Ours	0.945	0.957	0.485	0.686	0.949	0.956	0.494	0.687
832 × 624	Close-up	ViViD	0.931	0.946	0.456	0.673	0.933	0.948	0.460	0.676
		MagicTryon	0.940	0.935	0.465	0.648	0.942	0.935	0.470	0.653
		VACE	0.937	0.950	0.484	0.687	0.938	0.951	0.491	0.688
		Ours	0.939	0.957	0.483	0.670	0.940	0.957	0.494	0.675

Table 4: VBench evaluation results for the lower body category.

B.3 Upper body

Resolution	Type	Method	Unpaired		Paired			
			$VFID_R$↓	$VFID_I$↓	$VFID_R$↓	$VFID_I$↓	SSIM↑	LPIPS↓
1088 × 816	Full-shot	ViViD	0.368	9.058	0.326	7.711	0.877	0.122
		MagicTryon	**0.127**	**6.488**	0.110	**4.240**	0.912	0.089
		VACE	0.468	7.463	0.273	5.224	0.906	0.081
		Ours	0.172	6.738	**0.058**	4.546	**0.916**	**0.076**
832 × 624	Full-shot	ViViD	0.290	8.567	0.235	7.081	0.874	0.117
		MagicTryon	**0.101**	**6.474**	0.075	**4.286**	**0.914**	0.072
		VACE	0.602	7.803	0.363	5.453	0.896	0.088
		Ours	0.171	6.684	**0.057**	4.537	0.911	**0.069**
1088 × 816	Close-up	ViViD	1.296	9.095	1.026	8.082	0.774	0.197
		MagicTryon	0.765	**8.048**	0.595	6.595	**0.794**	0.170
		VACE	0.877	9.385	0.416	7.266	0.767	0.185
		Ours	**0.481**	8.248	**0.275**	**6.374**	0.791	**0.164**
832 × 624	Close-up	ViViD	0.888	8.657	0.673	7.441	0.764	0.187
		MagicTryon	0.603	**7.993**	0.514	6.470	**0.780**	0.159
		VACE	0.933	9.302	0.581	7.268	0.744	0.187
		Ours	**0.477**	8.164	**0.273**	**6.317**	0.774	**0.159**

Table 5: Quantitative comparison results for the upper body category.

Resolution	Type	Method	Unpaired				Paired			
			Subject Consistency	Background Consistency	Aesthetic Quality	Imaging Quality	Subject Consistency	Background Consistency	Aesthetic Quality	Imaging Quality
1088 × 816	Full-shot	ViViD	0.955	0.959	0.473	0.704	0.957	0.960	0.487	0.703
		MagicTryon	0.959	0.952	0.490	0.691	0.961	0.953	0.507	0.693
		VACE	0.957	0.963	0.498	0.709	0.958	0.964	0.515	0.708
		Ours	0.959	0.968	0.502	0.710	0.960	0.966	0.519	0.709
832 × 624	Full-shot	ViViD	0.956	0.963	0.470	0.699	0.958	0.964	0.483	0.698
		MagicTryon	0.960	0.968	0.491	0.699	0.962	0.968	0.504	0.701
		VACE	0.959	0.964	0.497	0.709	0.960	0.965	0.512	0.706
		Ours	0.959	0.965	0.502	0.690	0.959	0.965	0.516	0.692
1088 × 816	Close-up	ViViD	0.932	0.948	0.439	0.645	0.933	0.948	0.450	0.643
		MagicTryon	0.938	0.953	0.461	0.639	0.940	0.953	0.474	0.644
		VACE	0.937	0.951	0.474	0.675	0.937	0.951	0.480	0.674
		Ours	0.942	0.952	0.472	0.670	0.947	0.954	0.485	0.668
832 × 624	Close-up	ViViD	0.928	0.944	0.432	0.653	0.930	0.945	0.442	0.651
		MagicTryon	0.937	0.934	0.452	0.634	0.939	0.935	0.465	0.637
		VACE	0.933	0.949	0.471	0.668	0.935	0.950	0.482	0.672
		Ours	0.937	0.954	0.469	0.655	0.939	0.955	0.483	0.657

Table 6: VBench evaluation results for the upper body category.

B.4 Weighted average

To facilitate comparison, we report the weighted average results across the three categories. As shown in Table 7 and Table 8, both MagicTryon and our finetuned VACE model achieved the best results. While MagicTryon demonstrated superior performance on the $VFID_I$ and LPIPS metrics, our fine-tuned VACE model excelled on the other metrics, particularly in the paired setting. We attribute this advantage to the fine-tuning process and detailed garment image.

Resolution	Type	Method	Unpaired			Paired				
			$VFID_R$↓	$VFID_I$↓	VGID↑	$VFID_R$↓	$VFID_I$↓	VGID↑	SSIM↑	LPIPS↓
1088 × 816	Full-shot	ViViD	0.565	12.859	0.514	0.487	10.306	0.521	0.874	0.119
		MagicTryon	**0.187**	**9.783**	0.512	0.137	**6.168**	0.524	0.906	0.092
		VACE	0.454	11.260	0.522	0.282	7.290	0.525	0.899	0.083
		Ours	0.220	10.065	**0.524**	**0.076**	6.555	**0.534**	**0.910**	**0.078**
832 × 624	Full-shot	ViViD	0.389	12.194	0.506	0.285	9.211	0.508	0.871	0.111
		MagicTryon	**0.161**	9.865	0.520	0.106	**6.167**	0.514	**0.908**	0.074
		VACE	0.587	11.647	0.524	0.343	7.624	0.528	0.888	0.090
		Ours	0.219	**9.777**	**0.528**	**0.076**	6.520	**0.527**	0.905	**0.071**
1088 × 816	Close-up	ViViD	1.253	12.665	0.533	0.976	10.900	0.509	0.785	0.185
		MagicTryon	0.752	**11.285**	0.538	0.607	8.714	0.519	**0.807**	0.163
		VACE	0.913	12.844	0.533	0.429	9.331	0.515	0.778	0.174
		Ours	**0.535**	11.621	**0.540**	**0.274**	**8.429**	**0.524**	0.803	**0.154**
832 × 624	Close-up	ViViD	0.936	12.198	0.533	0.663	10.116	0.509	0.775	0.176
		MagicTryon	0.595	**11.262**	0.534	0.479	8.524	0.513	**0.791**	0.152
		VACE	1.039	12.783	0.545	0.593	9.572	0.519	0.755	0.178
		Ours	**0.530**	11.475	**0.548**	**0.273**	**8.349**	**0.526**	0.786	**0.148**

Table 7: Weighted quantitative comparison results across all categories.

Resolution	Type	Method	Unpaired				Paired			
			Subject Consistency	Background Consistency	Aesthetic Quality	Imaging Quality	Subject Consistency	Background Consistency	Aesthetic Quality	Imaging Quality
1088 × 816	Full-shot	ViViD	0.956	0.959	0.487	0.711	0.959	0.960	0.500	0.710
		MagicTryon	0.961	0.953	0.497	0.695	0.963	0.954	0.512	0.697
		VACE	0.958	0.964	0.507	0.714	0.960	0.965	0.523	0.713
		Ours	0.961	0.968	0.508	0.714	0.961	0.968	0.523	0.713
832 × 624	Full-shot	ViViD	0.958	0.963	0.478	0.705	0.960	0.964	0.491	0.705
		MagicTryon	0.962	0.969	0.496	0.703	0.964	0.969	0.508	0.704
		VACE	0.960	0.964	0.504	0.713	0.961	0.966	0.519	0.711
		Ours	0.961	0.966	0.507	0.697	0.961	0.966	0.522	0.697
1088 × 816	Close-up	ViViD	0.935	0.949	0.453	0.659	0.937	0.950	0.463	0.657
		MagicTryon	0.942	0.954	0.473	0.651	0.943	0.955	0.484	0.653
		VACE	0.940	0.952	0.488	0.684	0.940	0.952	0.493	0.681
		Ours	0.945	0.954	0.487	0.680	0.950	0.955	0.497	0.677
832 × 624	Close-up	ViViD	0.932	0.945	0.449	0.665	0.934	0.946	0.458	0.664
		MagicTryon	0.941	0.936	0.466	0.645	0.942	0.936	0.476	0.647
		VACE	0.937	0.950	0.487	0.679	0.939	0.951	0.496	0.680
		Ours	0.940	0.956	0.484	0.664	0.941	0.956	0.495	0.665

Table 8: Weighted VBench quantitative comparison results across all categories.

C. Additional qualitative results

Figure 7: Qualitative comparison for the dresses category.

Figure 8: Qualitative comparison for the lower body category.

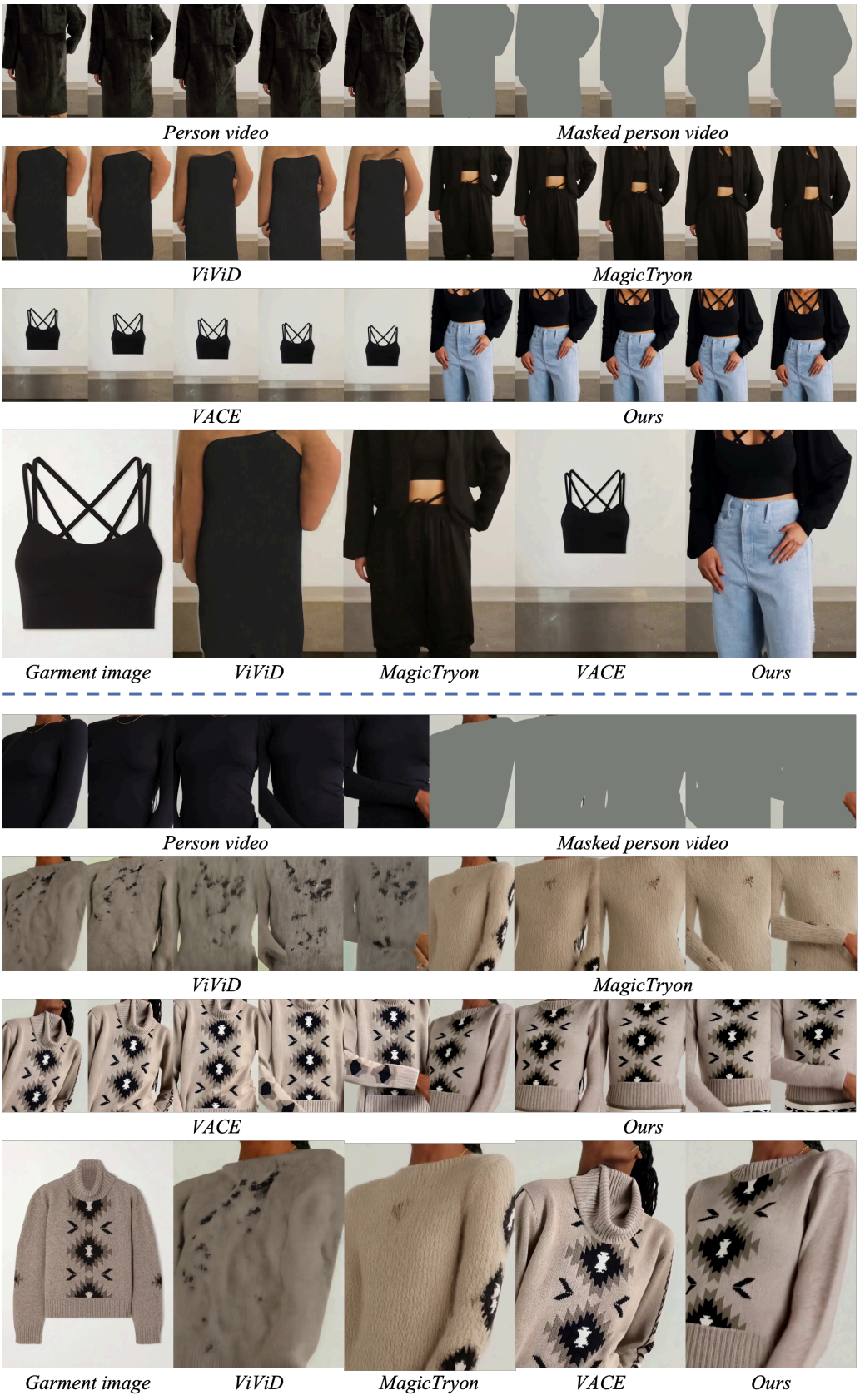

Figure 9: Qualitative comparison for the upper body category.

As illustrated in Figure 7,Figure 8,Figure 9, we compare the visual results of ViViD, MagicTryon, VACE, and our fine-tuned VACE across three distinct categories. To facilitate a detailed inspection, the images in the bottom row are single frames selected from the video results. For close-up videos, ViViD and MagicTryon often struggle to

accurately generate clear textures and patterns, whereas VACE tends to simply paste the original garment image onto the corresponding masked region. A consequence of this approach is that the inner collar label, which should be naturally occluded, often erroneously appears on the front. In contrast, our fine-tuned VACE not only achieves accurate try-on performance but also faithfully reconstructs the patterns of the input garment.

D. Ablation study

As shown in Table 9, Figure 10, Figure 11 and Figure 12, we conducted a series of comparative experiments based on the VACE model to better understand the impact of detailed garment images and the fine-tuning process.

The VACE model was not specifically trained or optimized for virtual try-on tasks, resulting in two distinct issues. First, the generated garments occasionally exhibit anomalous pure white artifacts. Second, VACE tends to naively paste the garment onto the masked region rather than simulating a realistic try-on, often leading to erroneous spatial relationships.

Simply inputting a detailed garment image during inference yields minimal improvement. While fine-tuning alone partially mitigates these issues, combining detailed garment inputs with fine-tuning enhances the model's understanding of the spatial relationship between the garment and the subject. This ultimately achieves a more natural and realistic try-on effect.

D.1 Quantitative results

Type	Category	LoRA Finetune	Detailed Image	Unpaired			Paired				
				$VFID_R$↓	$VFID_I$↓	VGID↑	$VFID_R$↓	$VFID_I$↓	VGID↑	SSIM↑	LPIPS↓
Full-shot	Dresses	✘	✘	0.565	17.928	0.522	0.488	12.585	0.525	0.855	0.107
		✘	✓	0.850	17.870	0.523	0.564	12.459	0.531	0.853	0.108
		✓	✘	**0.338**	17.452	0.520	0.186	11.735	0.529	0.867	0.101
		✓	✓	0.402	**17.033**	**0.524**	**0.149**	**11.604**	**0.534**	**0.869**	**0.099**
Full-shot	Lower body	✘	✘	0.316	12.184	–	0.093	6.128	–	0.929	0.063
		✘	✓	0.510	12.203	–	0.167	6.571	–	0.925	0.064
		✓	✘	**0.133**	9.792	–	**0.037**	5.343	–	0.939	0.061
		✓	✓	0.135	**9.752**	–	0.041	**5.525**	–	**0.939**	**0.059**
Full-shot	Upper body	✘	✘	0.468	7.463	–	0.273	5.224	–	0.906	0.081
		✘	✓	0.642	7.544	–	0.434	5.246	–	0.903	0.083
		✓	✘	0.215	6.817	–	0.075	4.578	–	0.915	0.077
		✓	✓	**0.172**	**6.738**	–	**0.058**	**4.546**	–	**0.916**	**0.076**
Full-shot	Weighted average	✘	✘	0.454	11.260	0.522	0.282	7.290	0.525	0.899	0.083
		✘	✓	0.661	11.290	0.523	0.400	7.381	0.531	0.896	0.084
		✓	✘	0.225	10.219	0.520	0.093	6.558	0.529	0.909	0.079
		✓	✓	**0.220**	**10.065**	**0.524**	**0.076**	**6.555**	**0.534**	**0.910**	**0.078**
Close-up	Dresses	✘	✘	1.217	17.385	0.533	0.431	11.538	0.515	0.774	0.172
		✘	✓	1.440	16.781	0.525	0.588	11.492	0.520	0.769	0.171
		✓	✘	0.818	15.824	0.529	**0.255**	**10.544**	0.517	**0.798**	**0.154**
		✓	✓	**0.731**	**15.710**	**0.540**	0.318	10.838	**0.524**	**0.798**	**0.154**
Close-up	Lower body	✘	✘	0.679	15.220	–	0.452	11.252	–	0.803	0.155
		✘	✓	0.895	15.291	–	0.727	11.145	–	0.795	0.156
		✓	✘	0.625	14.277	–	0.230	**10.039**	–	**0.831**	0.137
		✓	✓	**0.445**	**14.276**	–	**0.229**	10.131	–	**0.831**	**0.133**
Close-up	Upper body	✘	✘	0.877	9.385	–	0.416	7.266	–	0.767	0.185
		✘	✓	1.268	9.296	–	0.598	7.005	–	0.757	0.185
		✓	✘	0.494	8.463	–	**0.197**	6.551	–	0.790	0.166
		✓	✓	**0.481**	**8.248**	–	0.275	**6.374**	–	**0.791**	**0.164**
Full-shot	Weighted average	✘	✘	0.913	12.844	0.533	0.429	9.331	0.515	0.778	0.174
		✘	✓	1.218	12.666	0.525	0.628	9.162	0.520	0.770	0.174
		✓	✘	0.608	11.757	0.529	**0.220**	8.441	0.517	0.802	0.156
		✓	✓	**0.535**	**11.621**	**0.540**	0.274	**8.429**	**0.524**	**0.803**	**0.154**

Table 9: Ablation study of LoRA fine-tuning and detailed images across full-shot and close-up videos.

D.2 Qualitative results

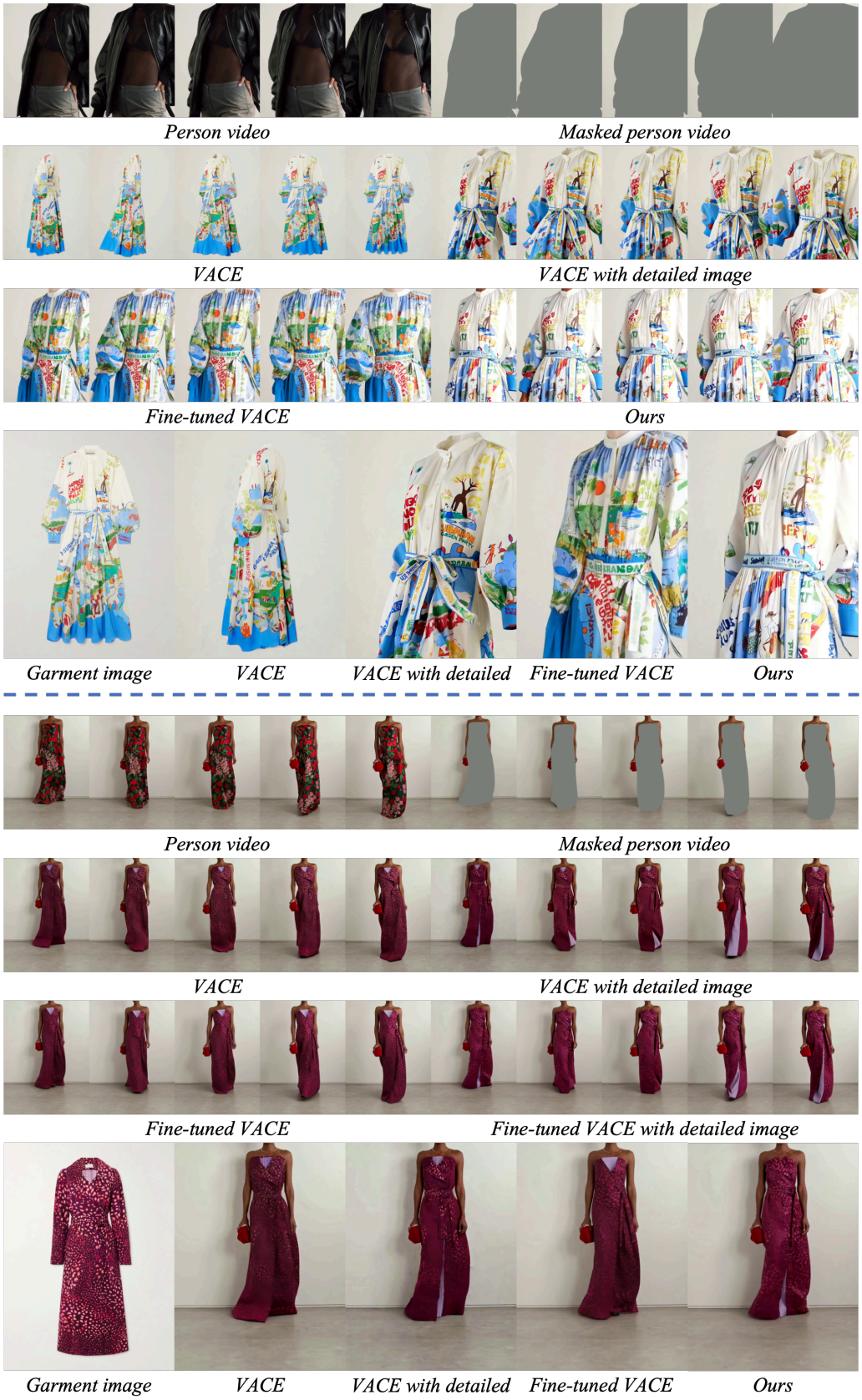

Figure 10: Qualitative comparison of the ablation study for the dresses category.

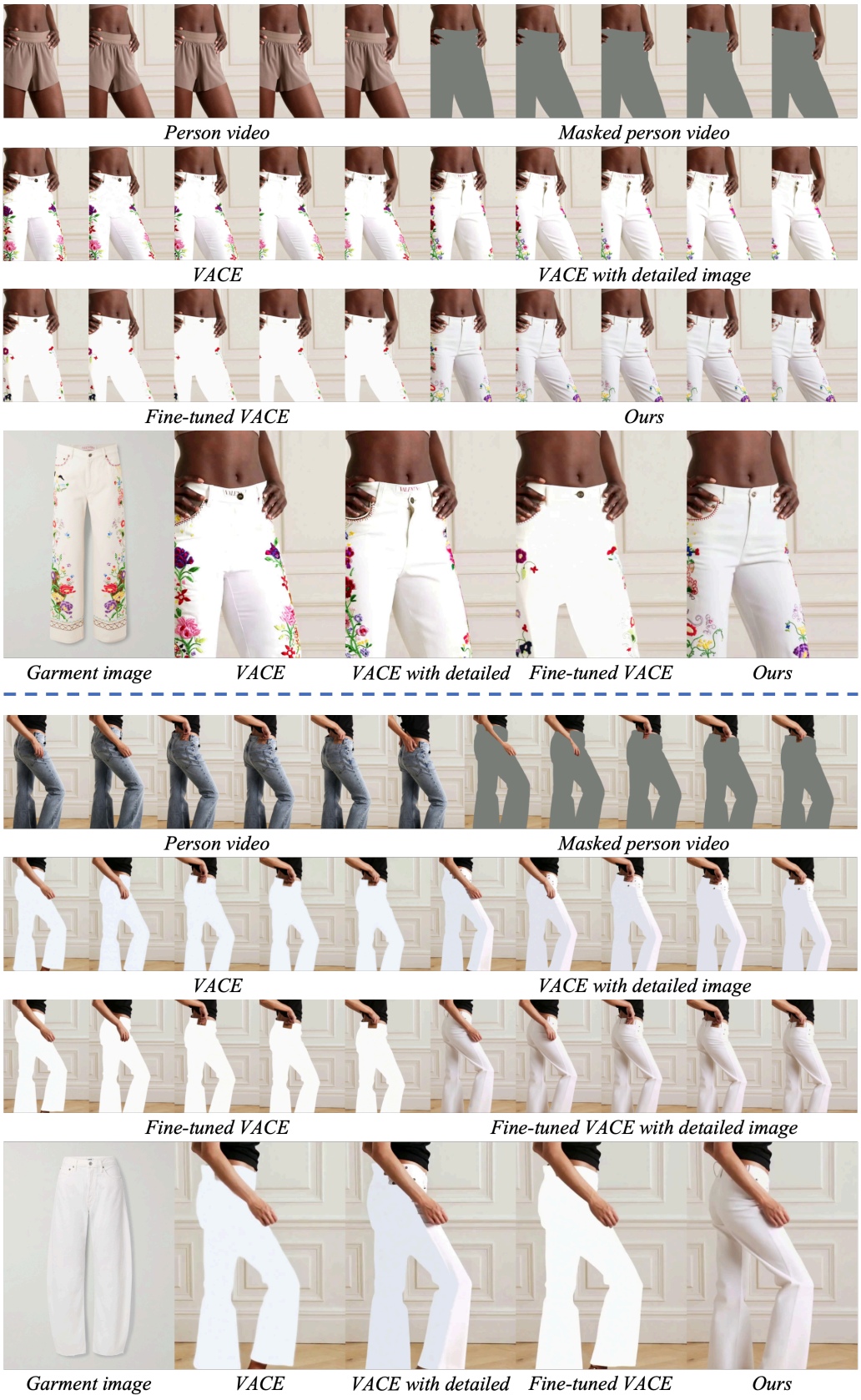

Figure 11: Qualitative comparison of the ablation study for the lower body category.

Figure 12: Qualitative comparison of the ablation study for the upper body category.

E. Potential applications

Compared to existing datasets for virtual try-on and related tasks, the Eevee dataset uniquely incorporates videos featuring diverse perspectives (full-shot and close-up) alongside comprehensive garment details. Consequently, it supports a broader spectrum of extended research directions:

High-Fidelity Image-based Virtual Try-on: Our dataset significantly advances image-based virtual try-on through two distinct approaches. First, by utilizing individual video frames as training data, we achieve a massive expansion in dataset scale. Eevee comprises 1,905,973 frames, an order of magnitude larger than the leading image-based dataset (DressCode), which contains only 161,376 images. Furthermore, unlike the ViViD dataset which is limited to a resolution of 832×624, whereas our video frames are provided at 1088×816, rivaling the quality of standard image-based benchmarks. Second, we support high-fidelity image-based virtual try-on by providing source images at 2400×1800. This resolution is more than four times higher than the standard 1024×768 found in datasets like VITON-HD, enabling the generation of results with superior texture fidelity and reduced blur.

Instruction-based Video Editing: One potential application of our dataset is the ability to perform instruction-based video editing, specifically enabling seamless transitions between full-shot and close-up perspectives. We can leverage Multimodal Large Language Models (MLLMs) to automatically generate descriptive instructions that guide the transformation from a full-shot person video to a close-up view, as well as the reverse operation. By compiling these full-shot videos, close-up videos, and their corresponding MLLM-generated text into aligned training pairs, we can train the model to understand and execute cinematic shot changes.

Fine-grained Texture Detail Generation: Our dataset offers significant potential for high-fidelity texture synthesis, addressing the common limitation where models fail to capture realistic details from single flat images. By providing detailed garment images paired with full in-shop garment images, we establish a unique resource for training models to "super-resolve" textures. This data alignment allows future research to treat the detailed garment image as ground truth, facilitating the development of specialized generative models.

F. Limitations

Dependency on Segmentation Masks: While image-based virtual try-on research has increasingly moved towards mask-free approaches to improve flexibility, our video-based method currently relies on explicit masking. The proposed framework functions as a masked video-to-video editing process , necessitating the use of external tools like OpenPose or Grounded SAM-2 to generate masks for the person and garment regions. This dependency introduces a bottleneck, as obtaining accurate human parsing results—particularly for the close-up videos introduced in our dataset—is often difficult and can lead to interference if the annotations are inaccurate.

High Computational Cost: Video-based virtual try-on demands significantly higher computational resources compared to static image tasks, especially when processing high-resolution data. Our experiments utilize the VACE model (built on Wan2.1) and processing videos at resolutions up to 1088×816. Consequently, the training and fine-tuning processes are computationally intensive. This high resource requirement presents a challenge for deploying such high-fidelity video try-on solutions in resource-constrained environments.