Visual-Textual Prompt Tuning for Unsupervised Multi-Domain Adaptation

Supplementary Material

Per-Pair Transfer Results

Table 3 provides the full per-pair classification accuracy across all 12 transfer directions on Office-Home, complementing the per-source averages reported in the main paper. Each column corresponds to a specific source-target domain pair, and results are grouped by source domain.

The trends observed in the main table are consistently reflected at the individual transfer direction level. VTPT with ViT-B/16 achieves the best accuracy in 11 out of 12 transfer directions, with the most pronounced improvements occurring in pairs where Clipart serves as the target domain. For instance, Ar→Cl improves from 51.6% under zero-shot CLIP to 70.6% under VTPT (+19.0%), and from 54.1% under DAPL to 70.6% (+16.5%). Similarly, Pr→Cl improves from 51.6% (CLIP) to 71.1% (VTPT, +19.5%), and Rw→Cl from 51.6% to 70.5% (+18.9%). These consistent gains across all Clipart-target directions confirm that the visual prompts are particularly effective at bridging the large appearance gap between photorealistic source domains and clipart-style imagery, which lacks the texture and photographic realism of natural images.

Transfer directions between photorealistic domains show a complementary pattern. Directions such as Pr→Rw (82.6% CLIP → 87.4% VTPT), Rw→Pr (81.9% → 90.7%), and Ar→Rw (82.6% → 90.8%) already achieve high accuracy under zero-shot CLIP, reflecting the smaller visual gap between these domains. VTPT still yields consistent improvements in these directions, suggesting that the visual prompts provide meaningful corrections even in low-gap transfer scenarios where the frozen image encoder already produces reasonably well-aligned features. Finally, the RN-101 variant of VTPT outperforms both CLIP and DAPL in all 12 directions without exception, further confirming that the gains stem from the coupled prompt learning framework rather than backbone capacity alone.

Per-Pair Ablation Results

Table 4 extends the ablation study from the main paper by reporting per-pair accuracy across all 12 transfer directions for each method variant. This allows a fine-grained analysis of where each component contributes most.

Comparing Text Prompting against VT Prompting across all backbones confirms that the visual prompt component yields consistent per-direction improvements, with gains most pronounced in Clipart-target directions. For RN-50, VT Prompting improves Ar→Cl from 54.55% to 54.98% and Ar→Rw from 83.96% to 85.38%, while the overall average rises from 73.95% to 74.24%. The gains become substantially larger with stronger backbones: with RN-101, VT Prompting achieves 78.45% overall versus 73.95% for the text-only baseline, and with ViT-B/16 the gap widens to 84.10%, a +10.15% absolute improvement that is distributed consistently across all 12 directions rather than concentrated in any particular subset.

The CSC comparison reveals a nuanced pattern at the per-direction level. While VT with CSC performs comparably to VT Prompting on most individual directions, it falls behind on average for both RN-101 (77.93% vs 78.45%) and ViT-B/16 (83.91% vs 84.10%). Inspecting individual directions shows that CSC tends to underperform most on Clipart-target directions. For example, VT with CSC (RN-101) scores 59.04% on Ar→Cl versus 59.34% for VT Prompting, and 58.90% on Pr→Cl versus 58.83%. This is consistent with the generalization argument in the main paper: CSC's per-class parameterization reduces the cross-domain regularization effect of joint multi-target training, and this effect is most visible in the highest-gap directions where generalization matters most.

Sensitivity Analysis: Confidence Threshold τ

Table 5 examines the sensitivity of VTPT to the pseudo-label confidence threshold τ using the ViT-B/16 backbone. We compare $\tau \in \{0.3, 0.5, 0.7\}$, where τ controls the minimum confidence required for a target sample to contribute to the unsupervised loss $\mathcal{L}_u$. $\tau = 0.5$ achieves the best overall average (84.10%) and is the most consistent across all 12 directions, confirming it as the optimal operating point. $\tau = 0.3$ yields a slightly lower average of 83.80%, suggesting that admitting lower-confidence pseudo-labels introduces noisy supervision that marginally degrades performance despite increasing the number of contributing samples. In particular, $\tau = 0.3$ underperforms on Clipart-target directions, where pseudo-label quality is more critical due to the larger distributional distance from photorealistic sources. $\tau = 0.7$ achieves an intermediate average of 83.99%, outperforming $\tau = 0.3$ but falling short of $\tau = 0.5$. This is consistent with the expected trade-off: a stricter threshold improves pseudo-label precision but reduces the effective number of contributing target samples, limiting adaptation particularly in low-confidence target domains. Notably, $\tau = 0.7$ leads on Product-target and Real-World-to-Product directions, where high-confidence predictions are more abundant, but lags on directions involving Clipart as target, where the domain gap restricts the pool of reliable pseudo-labels.

Table 3. Per-pair classification accuracy (%) on Office-Home across all 12 transfer directions. $*$: single-target method applied independently per target domain. **Bold**: best result per column.

Method	**Backbone**	**Ar** $\rightarrow$			**Cl** $\rightarrow$			**Pr** $\rightarrow$			**Rw** $\rightarrow$			**Avg**
		Cl	Pr	Rw	Ar	Pr	Rw	Ar	Cl	Rw	Ar	Cl	Pr	
CLIP* [28]	RN-50	51.6	81.9	82.6	71.9	81.9	82.6	71.9	51.6	82.6	71.9	51.6	81.9	72.0
DAPL* [6]	RN-50	54.1	84.3	84.8	74.4	83.7	85.0	74.5	54.6	84.8	75.2	54.7	83.8	74.5
VTPT (Ours)	RN-101	59.3	87.5	87.1	80.0	87.9	87.5	80.1	58.8	87.4	78.4	59.4	88.1	78.5
VTPT (Ours)	ViT-B/16	**70.6**	**90.7**	**90.8**	**84.1**	**90.9**	**90.8**	**83.9**	**71.1**	**91.0**	**84.3**	**70.5**	**90.7**	**84.1**

Ar: Art, **Cl**: Clipart, **Pr**: Product, **Rw**: Real-World.

Table 4. Per-pair ablation accuracy (%) on Office-Home across all method variants and backbone configurations. **Bold**: best result per column.

Method	**Ar→Cl**	**Ar→Pr**	**Ar→Rw**	**Cl→Ar**	**Cl→Pr**	**Cl→Rw**	**Pr→Ar**	**Pr→Cl**	**Pr→Rw**	**Rw→Ar**	**Rw→Cl**	**Rw→Pr**	**Avg**
Text Prompting (RN-50)	54.55	82.95	83.96	73.88	83.55	84.14	73.22	54.80	84.94	73.71	54.48	83.19	73.95
VT Prompting (RN-50)	54.98	83.74	85.38	72.48	83.87	84.69	74.00	54.73	84.71	73.55	55.03	83.69	74.24
VT with CSC (RN-50)	55.05	84.57	85.61	73.92	83.17	84.76	73.47	53.86	84.62	73.42	54.32	82.29	74.09
VT Prompting (RN-101)	59.34	87.47	87.10	80.02	87.86	87.45	80.14	58.83	87.38	78.37	59.36	88.08	78.45
VT with CSC (RN-101)	59.04	86.46	86.41	78.78	87.63	87.84	78.57	58.90	86.55	79.52	59.40	86.01	77.93
VT Prompting (ViT-B/16)	**70.63**	**90.67**	90.75	**84.10**	**90.88**	**90.73**	83.93	**71.11**	**90.96**	84.34	**70.47**	**90.67**	**84.10**
VT with CSC (ViT-B/16)	69.99	90.45	**90.88**	**84.10**	90.63	90.43	**84.10**	70.77	90.45	**84.67**	70.19	90.27	83.91

Ar: Art, **Cl**: Clipart, **Pr**: Product, **Rw**: Real-World.

Table 5. Sensitivity of VTPT (ViT-B/16) to the pseudo-label confidence threshold τ. Per-pair accuracy (%) across all 12 transfer directions on Office-Home. **Bold**: best result per column.

Method	**Ar→Cl**	**Ar→Pr**	**Ar→Rw**	**Cl→Ar**	**Cl→Pr**	**Cl→Rw**	**Pr→Ar**	**Pr→Cl**	**Pr→Rw**	**Rw→Ar**	**Rw→Cl**	**Rw→Pr**	**Avg**
VT Prompting τ=0.3	69.99	90.67	90.64	**84.10**	91.12	**90.93**	83.23	70.42	**91.07**	83.77	69.64	90.74	83.80
VT Prompting τ=0.5	**70.63**	90.67	**90.75**	**84.10**	90.88	90.73	**83.93**	**71.11**	90.96	**84.34**	70.47	90.67	**84.10**
VT Prompting τ=0.7	70.58	**91.01**	90.08	83.56	**91.35**	90.64	83.60	70.86	90.66	83.81	**70.70**	**91.06**	83.99

Ar: Art, **Cl**: Clipart, **Pr**: Product, **Rw**: Real-World.