A. Proof of Theorem 1

Its Lagrangian function can be written as

$$\mathcal{L}(\boldsymbol{d}, v, \boldsymbol{\xi}, \boldsymbol{\eta}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = v + C_1 \cdot \sum_{j \in \mathcal{I}_a} \xi_j + C_2 \cdot \sum_{i \in \mathcal{I}_p} \sum_{j \in \mathcal{I}_a} \eta_{ij} + \frac{1}{2}\|\boldsymbol{d}\|^2 + \sum_{i \in \mathcal{I}_p} \alpha_i (\langle \nabla_{\boldsymbol{\theta}} \ell_m(\boldsymbol{\theta}^{(\tau)}), \boldsymbol{d} \rangle - v)$$
$$+ \sum_{j \in \mathcal{I}_a} \alpha_j (\langle \nabla_{\boldsymbol{\theta}} \ell_j(\boldsymbol{\theta}^{(\tau)}), \boldsymbol{d} \rangle - v - \xi_j) + \sum_{i \in \mathcal{I}_p} \sum_{j \in \mathcal{I}_a} \beta_{ij} (\langle \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^{(\tau)}), \boldsymbol{d} \rangle - \langle \nabla_{\boldsymbol{\theta}} \ell_j(\boldsymbol{\theta}^{(\tau)}), \boldsymbol{d} \rangle - \eta_{ij}) - \sum_{i \in \mathcal{I}_a} \mu_i \xi_i - \sum_{i \in \mathcal{I}_p} \sum_{j \in \mathcal{I}_a} \lambda_{ij} \eta_{ij}, \tag{18}$$

where α_m and β_m are Lagrange multipliers. Then we have

$$\begin{aligned}
&\frac{\partial \mathcal{L}}{\partial \boldsymbol{d}} = \boldsymbol{d} + \sum_{m=1}^{M} \alpha_m \nabla_{\boldsymbol{\theta}} \ell_m(\boldsymbol{\theta}^{(\tau)}) + \sum_{i \in \mathcal{I}_p} \sum_{j \in \mathcal{I}_a} \beta_{ij} (\nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^{(\tau)}) - \nabla_{\boldsymbol{\theta}} \ell_j(\boldsymbol{\theta}^{(\tau)})) = 0, \\
\rightarrow \quad &\boldsymbol{d} = -\sum_{m=1}^{M} \alpha_m \nabla_{\boldsymbol{\theta}} \ell_m(\boldsymbol{\theta}^{(\tau)}) - \sum_{i \in \mathcal{I}_p} \sum_{j \in \mathcal{I}_a} \beta_{ij} (\nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^{(\tau)}) - \nabla_{\boldsymbol{\theta}} \ell_j(\boldsymbol{\theta}^{(\tau)})), \\
&\frac{\partial \mathcal{L}}{\partial v} = 1 - \sum_{m=1}^{M} \alpha_m = 0, \quad \rightarrow \quad \sum_{m=1}^{M} \alpha_m = 1, \\
&\frac{\partial \mathcal{L}}{\partial \xi_j} = C_1 - \alpha_j - \mu_j = 0, \quad \rightarrow \quad \alpha_j + \mu_j = C_1, \ \forall j \in \mathcal{I}_a \\
&\frac{\partial \mathcal{L}}{\partial \eta_{ij}} = C_2 - \beta_{ij} - \lambda_{ij} = 0, \quad \rightarrow \quad \beta_{ij} + \lambda_{ij} = C_2, \forall i \in \mathcal{I}_p, \forall j \in \mathcal{I}_a.
\end{aligned} \tag{19}$$

From the Karush–Kuhn–Tucker (KKT) conditions, we have

$$\boldsymbol{d}^* + \sum_{m=1}^{M} \alpha_m \nabla_{\boldsymbol{\theta}} \ell_m(\boldsymbol{\theta}^{(\tau)}) + \sum_{i \in \mathcal{I}_p} \sum_{j \in \mathcal{I}_a} \beta_{ij} (\nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^{(\tau)}) - \nabla_{\boldsymbol{\theta}} \ell_j(\boldsymbol{\theta}^{(\tau)})) = 0, \tag{20}$$

and for each $i \in \mathcal{I}_p, j \in \mathcal{I}_a$ $m \in \mathcal{I}_p, \mathcal{I}_a$, we have

$$\begin{aligned}
&\langle \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^{(\tau)}), \boldsymbol{d}^* \rangle \leq v^*, \ \langle \nabla_{\boldsymbol{\theta}} \ell_j(\boldsymbol{\theta}^{(\tau)}), \boldsymbol{d}^* \rangle \leq v^* + \xi_j^*, \\
&\langle \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^{(\tau)}), \boldsymbol{d}^* \rangle \leq \langle \nabla_{\boldsymbol{\theta}} \ell_j(\boldsymbol{\theta}^{(\tau)}), \boldsymbol{d}^* \rangle + \eta_{ij}^*, \\
&\alpha_m \geq 0, \ \xi_j^* \geq 0, \ \eta_{ij}^* \geq 0, \ \beta_{ij} \geq 0, \ \mu_j \geq 0, \ \lambda_{ij} \geq 0, \\
&\alpha_i (\langle \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^{(\tau)}), \boldsymbol{d}^* \rangle - v^*) = 0, \ \alpha_j (\langle \nabla_{\boldsymbol{\theta}} \ell_j(\boldsymbol{\theta}^{(\tau)}), \boldsymbol{d}^* \rangle - v^* - \xi_j^*) = 0, \\
&\beta_{ij} (\langle \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^{(\tau)}), \boldsymbol{d}^* \rangle - \langle \nabla_{\boldsymbol{\theta}} \ell_j(\boldsymbol{\theta}^{(\tau)}), \boldsymbol{d}^* \rangle - \eta_{ij}^*) = 0, \ \mu_j \xi_j^* = 0, \ \lambda_{ij} \eta_{ij}^* = 0,
\end{aligned} \tag{21}$$

1. if $\boldsymbol{d}^* = \boldsymbol{0}$, then $\langle \nabla_{\boldsymbol{\theta}} \ell_m(\boldsymbol{\theta}^{(\tau)}), \boldsymbol{d}^* \rangle = 0, \forall m$, and immediately satisfies the conclusion.
2. if $\boldsymbol{d}^* \neq \boldsymbol{0}$, then we have

$$v^* = -\|\boldsymbol{d}\|^2 - \sum_{j \in \mathcal{I}_a} \alpha_j \xi_j^* - \sum_{i \in \mathcal{I}_p} \sum_{j \in \mathcal{I}_a} \beta_{ij} \eta_{ij}^*.$$

Thus

$$\begin{aligned}
&\forall i \in \mathcal{I}_p, \langle \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^{(\tau)}), \boldsymbol{d}^* \rangle \leq v^* = -\|\boldsymbol{d}\|^2 - C_1 \sum_{j \in \mathcal{I}_a} \xi_j^* - C_2 \sum_{i \in \mathcal{I}_p} \sum_{j \in \mathcal{I}_a} \eta_{ij}^* \\
&\forall j \in \mathcal{I}_a, \langle \nabla_{\boldsymbol{\theta}} \ell_j(\boldsymbol{\theta}^{(\tau)}), \boldsymbol{d}^* \rangle \leq v^* + \xi_j^* = -\|\boldsymbol{d}\|^2 - C_1 \sum_{j \in \mathcal{I}_a} \xi_j^* - C_2 \sum_{i \in \mathcal{I}_p} \sum_{j \in \mathcal{I}_a} \eta_{ij}^* + \xi_j^*,
\end{aligned} \tag{22}$$

which means that for the main tasks and the auxiliary task satisfying $-\|\boldsymbol{d}\|^2 - C_1 \sum_{j \in \mathcal{I}_a} \xi_j^* - C_2 \sum_{i \in \mathcal{I}_p} \sum_{j \in \mathcal{I}_a} \eta_{ij}^* + \xi_j^* \leq 0$, their loss function values will decrease.

B. Details of Hybrid Solver for Eq. (14)

Denote all the integer from a to b with $[a..b]$. The dual problem to solve is

$$\begin{aligned}
&\min_{\boldsymbol{x}} \boldsymbol{x}^T Q \boldsymbol{x} \ , \boldsymbol{x} \in \mathbb{R}^{M_t+M_p+M_t*M_p} \\
&\text{s.t. } x_i \geq 0, i \in [1..M_t+M_p],\ x_i \leq C_1, i \in [M_t..M_t+M_p],\ \sum_{i\in[1..M_t]} x_i = 1, \\
&\quad 0 \leq x_j \leq C_2, j \in [M_t+M_p+1..M_t+M_p+M_t*M_p], \\
&\quad \sum_{i\in\Pi_j} x_i \leq x_j, j \in [M_t+1..M_t+M_p],\ \Pi_j = \{M_t+M_p+M_p*k+j\}, k \in [1..M_t],
\end{aligned} \tag{23}$$

with Q as $\begin{pmatrix} [\Delta_t,\Delta_p]^T[\Delta_t,\Delta_p] & [\Delta_t,\Delta_p]^T(\Delta_t A-\Delta_p P^T) \\ (\Delta_t A-\Delta_p P^T)^T[\Delta_t,\Delta_p] & (\Delta_t A-\Delta_p P^T)^T(\Delta_t A-\Delta_p P^T) \end{pmatrix}$. When updating $x_i, i \in [1..M_t+M_p]$, we choose a working set B with size $|B|=2$, and solve the following subproblem.

$$\begin{aligned}
&\min_{x_i,x_j} Q_{ii}x_i^2 + Q_{jj}x_j^2 + 2Q_{ij}x_ix_j + 2\sum_k (Q_{ik}x_i + Q_{jk}x_j)x_k,\ k \in [1..M_t+M_p+M_t*M_p], k \neq i,j, \\
&\text{s.t. } x_i + x_j = 1 - \sum_k x_k, k \in [1..M_t+M_p], k \neq i,j, \\
&\quad \sum_{l\in\Pi_i} x_l \leq x_i \leq C_1, \forall i \in [M_t+1..M_t+M_p],\ 0 \leq x_i \leq C_1, \forall i \in [1..M_t].
\end{aligned} \tag{24}$$

Assume Q_{ii} and Q_{jj} are nonzero, then this problem is reduced to a single-variable quadratic optimization with bound constraint. As established previously, appropriately choosing the working set B at each iteration can speed up the convergence of the overall optimization. One of the heuristics is to find the most violating pair according to the KKT condition. And the KKT condition for $x_i, i \in [1..M_t+M_p]$ is

$$\begin{aligned}
&(Q\boldsymbol{x})_i + \mu_i - \nu_i + \lambda = 0,\ \sum_i x_i = 1, i \in [1..M_t+M_p], \\
&L \leq x_i \leq C_1,\ \mu_i \geq 0,\ \nu_i \geq 0,\ \mu_i(x_i - C_1) = 0,\ \nu_i(x_i - L) = 0,
\end{aligned} \tag{25}$$

where L is the new lower bound of x_i, either 0 or $\sum_{l\in\Pi_i} x_l$. For $i \in \mathcal{I}_{low} = \{i | x_i > L\}$, we have $\lambda \leq -(Q\boldsymbol{x})_i$, whereas for $i \in \mathcal{I}_{up} = \{i | x_i < C_1\}$, we have $\lambda \geq -(Q\boldsymbol{x})_i$. Therefore, $\max\{-(Q\boldsymbol{x})_i, i \in \mathcal{I}_{up}\}\} \leq \min\{-(Q\boldsymbol{x})_i, i \in \mathcal{I}_{low}\}\}$. According to this requirement, an index pair (i,j) is called a maximal violating pair if

$$i = \text{argmax}_t\{-(Q\boldsymbol{x})_t | t \in \mathcal{I}_{low}\},\ j = \text{argmin}_t\{-(Q\boldsymbol{x})_t | t \in \mathcal{I}_{up}.\} \tag{26}$$

For $i \in [M_t+M_p+1..M_t+M_p+M_t*M_p]$, we consider a random selection of a single variable to descent each time. And the subproblem is thus

$$\min_{x_i}\ Q_{ii}x_i^2 + 2\sum_{k\in\mathcal{I}_1, k\neq i} Q_{ik}x_ix_k,\ \text{s.t. } 0 \leq x_i \leq \min(C_2, x_k - \sum_{l\in\Omega_i} x_l) \tag{27}$$

The algorithm for "HybridSolver" is shown as Algorithm 2.

Algorithm 2: HybridSolver

Let $\boldsymbol{x}$ be a feasible point and calculate $\boldsymbol{q}$, `outer_step` = 0;
while $\boldsymbol{x}$ *is not optimal* **do**
  Calculate the full gradient $\nabla_{\boldsymbol{x}} f(\boldsymbol{x})$;
  if `outer_step` *is even* **then**
    Update $x_i, x_j, i \in \mathcal{I}_1$; ▷ CD with linear constraints
    Determine the working set by (26) ;
    for $s \leftarrow 1$ **to** r **do**
      $(i, j) \leftarrow (i_s, j_s)$;
      Calculate $(Q\boldsymbol{x})_i$ and $(Q\boldsymbol{x})_j$;
      if (i, j) *is not violating pair* **then**
      else
        Solve the subproblem (24)
      end
    end
  else
    Update $x_i, i \in \mathcal{I}_2$; ▷ CD without a linear constraint
    Random permute $\{1, \ldots, l\}$ to $\{\pi(1), \ldots, \pi(l)\}$;
    for $s \leftarrow 1$ **to** l **do**
      $i \leftarrow \pi(s)$;
      Solve the subproblem (27) ;
    end
  end
  `outer_step` = `outer_step` + 1;
end

C. Detailed Results on CelebA and CIFAR-100 Datasets in Figure 5

Table 3: CelebA dataset multi-label classification error per attribute for all algorithms.

	Single Task	tMOO-MTL	LinSca-lar	Uncer-tainty	MOO-MTL	Ours		Single Task	tMOO-MTL	LinSca-lar	Uncer-tainty	MOO-MTL	Ours
Attr. 0	7.16	6.95	7.11	7.18	6.17	**5.99**	Attr. 20	1.61	/	1.61	1.58	1.43	**1.18**
Attr. 1	**14.38**	17.78	17.30	16.77	14.87	14.5	Attr. 21	6.20	/	7.18	7.73	6.26	**6.06**
Attr. 2	19.25	20.49	20.99	20.56	**18.35**	18.41	Attr. 22	4.14	/	4.38	4.08	3.81	4.13
Attr. 3	16.79	17.3	17.82	18.45	16.06	**15.54**	Attr. 23	**6.57**	/	8.32	8.80	6.47	6.63
Attr. 4	1.20	1.32	1.25	1.17	**1.08**	1.21	Attr. 24	5.38	/	5.01	5.12	4.23	**4.16**
Attr. 5	4.75	4.74	4.91	4.95	4.13	**4.08**	Attr. 25	24.82	/	27.59	26.94	23.87	**23.73**
Attr. 6	14.24	14.39	20.97	15.17	**14.08**	14.75	Attr. 26	3.40	/	3.54	3.78	**3.16**	3.2
Attr. 7	17.74	18.03	18.53	18.84	17.25	**17.21**	Attr. 27	22.74	/	26.74	26.21	22.45	**21.94**
Attr. 8	8.87	10.21	10.22	10.19	8.42	**8.29**	Attr. 28	5.82	/	6.14	6.17	5.16	**5.05**
Attr. 9	5.09	5.27	5.29	5.44	4.60	**4.57**	Attr. 29	5.18	/	5.55	5.40	**4.87**	4.95
Attr. 10	4.02	5.92	4.14	4.33	3.60	**3.45**	Attr. 30	3.79	/	3.29	3.24	3.03	**2.97**
Attr. 11	15.34	15.52	16.22	16.64	**14.56**	14.69	Attr. 31	7.18	/	8.05	8.40	6.92	**6.64**
Attr. 12	7.68	14.25	8.42	8.85	7.41	**7.12**	Attr. 32	17.25	/	18.21	18.15	15.93	**15.38**
Attr. 13	5.15	6.12	5.17	5.26	**4.52**	4.56	Attr. 33	15.55	/	16.53	16.19	13.80	**13.76**
Attr. 14	4.13	4.91	4.14	4.17	3.54	**3.37**	Attr. 34	9.76	/	11.12	11.46	9.73	**9.29**
Attr. 15	0.52	6.96	0.81	0.62	0.56	**0.48**	Attr. 35	1.13	/	1.15	1.08	1.08	**1.03**
Attr. 16	3.94	3.91	4.00	3.99	**3.46**	3.62	Attr. 36	7.56	/	7.91	8.06	**7.18**	7.34
Attr. 17	2.66	4.87	2.39	2.35	2.16	**2.14**	Attr. 37	11.90	/	13.27	13.47	11.19	**11.01**
Attr. 18	9.01	11.4	8.79	8.84	7.83	**7.75**	Attr. 38	**3.29**	/	3.80	4.04	3.51	3.33
Attr. 19	12.27	24.64	13.78	13.86	**11.29**	11.79	Attr. 39	13.40	/	13.25	13.78	11.95	**11.82**

Table 4: CIFAR-100 dataset 5-way multi-label classification error per attribute for all algorithms.

	tMOO-MTL	LinScalar	Uncertainty	MOO-MTL	Ours
Task 0	24.00	**21.29**	26.87	26.83	21.60
Task 1	20.40	20.07	21.26	21.01	**17.40**
Task 2	19.00	23.05	22.65	25.96	**18.40**
Task 3	**21.80**	25.33	25.06	25.33	23.20
Task 4	15.80	13.65	**12.86**	13.39	15.00
Task 5	23.00	24.09	26.53	25.96	**17.40**
Task 6	**13.20**	18.30	22.16	19.76	**13.20**
Task 7	25.20	27.04	**20.99**	23.71	21.40
Task 8	15.40	24.09	24.03	22.64	**13.20**
Task 9	6.80	16.91	14.44	14.95	**5.80**
Task 10	-	15.46	15.05	**12.36**	12.80
Task 11	-	21.04	24.38	**18.56**	22.20
Task 12	-	16.93	**12.21**	**12.21**	19.20
Task 13	-	**22.85**	**22.85**	23.81	23.40
Task 14	-	**19.64**	20.42	17.55	46.80
Task 15	-	**20.92**	24.40	24.48	27.60
Task 16	-	20.24	19.56	16.57	**19.40**
Task 17	-	23.35	**22.55**	23.45	24.80
Task 18	-	15.42	16.93	13.01	**7.80**
Task 19	-	18.65	16.93	17.71	**10.80**