

Towards Calibrated Gradient-based Multi-Task Learning

Supplementary Material

7. Appendix

7.1. A. Experimental Details

7.1.1. A.1 Dataset and Task Descriptions

We evaluate VarGrad across three supervised MTL scenarios—large-scale classification, regression, and dense prediction—each presenting unique optimization challenges.

Image-Level Multi-Task Classification. CelebA [23] is a large-scale face attribute dataset containing over 200K celebrity images, each annotated with 40 binary attributes (e.g., Smiling, Wavy Hair, Mustache). This benchmark is framed as a 40-task binary classification problem and is commonly used to evaluate the scalability of MTL methods.

Multi-Task Regression. QM9 [29] consists of 134K organic molecules represented as graphs. Each molecule is annotated with 11 real-valued regression targets corresponding to quantum chemical properties with different numerical scales. This task is used to test the ability of MTL methods to handle regression with varying magnitudes and distributions.

Dense Prediction. NYU-v2 [32] includes 1,449 indoor RGB-D images with three pixel-wise tasks: 13-class semantic segmentation, depth estimation, and surface normal prediction. Cityscapes [6] contains 5,000 urban street scenes with two pixel-wise tasks: 7-class segmentation and depth estimation. These datasets evaluate the robustness of MTL methods on dense prediction with heterogeneous task types and output structures.

7.1.2. A.2 Training Details

We closely follow experimental settings from prior work [22, 28], with minor adjustments for consistency.

- CelebA: A 9-layer CNN is used as the shared backbone, with a separate linear classification head for each attribute. Models are trained using Adam optimizer with an initial learning rate of 3×10^{-4} , batch size 256, and for 15 epochs.
- QM9: We adopt the official PyTorch Geometric implementation [11], using the standard 110K/10K/10K split for training, validation, and testing. Models are trained for 300 epochs with batch size 120. The initial learning rate is set to 1×10^{-3} , and a scheduler reduces the learning rate upon validation stagnation.

- NYU-v2: We use the MTAN architecture [26] based on SegNet [1], with task-specific attention modules. Models are trained for 200 epochs with a batch size of 2. The learning rate is 1×10^{-4} for the first 100 epochs and halved afterward.
- Cityscapes: We use the same architecture and training schedule as for NYU-v2, except with a batch size of 8 due to higher image resolution. The total training duration is also 200 epochs.

All experiments are conducted using PyTorch 2.7.1 and CUDA 12.4 on NVIDIA A40 GPUs (48GB). For each method and dataset, we report results averaged over three runs with different random seeds. This setting ensures statistical robustness and reduces the impact of randomness on performance metrics. We apply the same random seeds across all methods—including our method and all baseline approaches such as MGDA-R, FairGrad-R, to ensure a fair comparison under identical training conditions.

7.1.3. A.3 Hyperparameter Selection

Our method involves two primary hyperparameters: the variance control coefficient β and the selective update threshold τ . We perform empirical tuning for both based on validation performance.

Variance Coefficient β . The hyperparameter $\beta \in (0, 1)$ controls the strength of variance reduction in the proposed correction term. We perform grid search over the range $[0.75, 0.95]$ with an interval of 0.025, and observe consistent stability and convergence benefits within this interval. A value of $\beta = 0.85$ is used across all experiments for simplicity and generalization. We find that β values below 0.75 tend to under-compensate for variance, while values above 0.95 may overly smooth gradient dynamics, leading to slower convergence in some tasks.

Imbalance Threshold τ . The threshold $\tau > 1$ governs when to activate multi-task gradient coordination. Specifically, we compute the inter-task imbalance ratio

$$r = \frac{\max_t \omega_t}{\min_t \omega_t}, \quad (17)$$

where ω_t measures per-task weights. When $r > \tau$, we invoke a multi-task optimizer (e.g., FairGrad); otherwise, we fall back to unweighted loss summation to reduce unnecessary coordination overhead. We fix $\tau = 1.5$ in all experiments. This value offers a good trade-off between responsiveness and stability: it ensures that coordination is

triggered only under meaningful imbalance, avoiding fluctuations from noisy updates.

7.2. B. Theoretical Analysis

We analyze the convergence of VarGrad. Let the weighted objective be defined as:

$$F(\mathbf{x}) := \sum_{t=1}^T w_t F_t(\mathbf{x}) \quad (18)$$

where each F_t is a task-specific loss function and $w_t \geq 0$ are fixed weights such that $\sum_{t=1}^T w_t = 1$.

To proceed, we introduce a set of assumptions characterizing the structure of each task loss and the behavior of the stabilized gradient estimators. These assumptions ensure the landscape is smooth and that the gradient bias and variance diminish in a controlled manner over time.

Assumption 1 (Smoothness) *Each task loss function F_t is differentiable and L -smooth, i.e., for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,*

$$F_t(\mathbf{y}) \leq F_t(\mathbf{x}) + \langle \nabla F_t(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|^2. \quad (19)$$

This standard smoothness condition ensures that the loss landscape behaves regularly, enabling us to quantify the descent progress made by gradient-based updates. Next, we turn our attention to the properties of the stabilized gradient estimators used by VarGrad. Since these estimators are noisy and potentially biased due to task-specific stabilization, we impose the following lemmas on the expected bias.

Lemma 1 (Controlled Bias) *Let m_t^k denote the momentum-based gradient estimator for task t at iteration k . Then there exists a bias term $\bar{\delta}_t^k$ such that:*

$$\mathbb{E}[m_t^k] = \mathbb{E}[\nabla F_t(x_k)] + \bar{\delta}_t^k, \quad \text{with} \quad \|\bar{\delta}_t^k\| \leq C_1 k^{1/3}. \quad (20)$$

Proof of Lemma 1.

Definitions and Assumptions.

Momentum Estimator:

$$m_t^k = (1 - \beta_k) (m_t^{k-1} + \nabla f_t(x_k, \xi_k) - \nabla f_t(x_{k-1}, \xi_k)) + \beta_k \nabla f_t(x_k, \xi_k) \quad (21)$$

where $\nabla f_t(x, \xi)$ is a stochastic gradient such that

$$\mathbb{E}_\xi[\nabla f_t(x, \xi)] = \nabla F_t(x). \quad (22)$$

Bias Definition:

$$\bar{\delta}_t^k := \mathbb{E}[m_t^k] - \mathbb{E}[\nabla F_t(x_k)] \quad (23)$$

Lipschitz Gradient: The true gradient ∇F_t is L -Lipschitz:

$$\|\nabla F_t(x) - \nabla F_t(y)\| \leq L \|x - y\|, \quad \forall x, y. \quad (24)$$

Bounded Step: There exists a constant G such that:

$$\mathbb{E}[\|x_k - x_{k-1}\|] \leq \eta_{k-1} G. \quad (25)$$

Recursive Inequality for Bias Norm. Taking expectation on both sides of the update rule:

$$\begin{aligned} \mathbb{E}[m_t^k] &= (1 - \beta_k) (\mathbb{E}[m_t^{k-1}] + \mathbb{E}[\nabla F_t(x_k)] - \mathbb{E}[\nabla F_t(x_{k-1})]) \\ &\quad + \beta_k \mathbb{E}[\nabla F_t(x_k)] \end{aligned} \quad (26)$$

Then,

$$\begin{aligned} \bar{\delta}_t^k &= \mathbb{E}[m_t^k] - \mathbb{E}[\nabla F_t(x_k)] \\ &= (1 - \beta_k) (\mathbb{E}[m_t^{k-1}] - \mathbb{E}[\nabla F_t(x_{k-1})]) = (1 - \beta_k) \bar{\delta}_t^{k-1} \end{aligned} \quad (27)$$

To account for the error introduced by $x_k \neq x_{k-1}$ following the approach of [8], we use a tighter inequality:

$$\begin{aligned} \|\bar{\delta}_t^k\| &\leq (1 - \beta_k) \|\bar{\delta}_t^{k-1}\| + \beta_k \|\mathbb{E}[\nabla F_t(x_k) - \nabla F_t(x_{k-1})]\| \\ &\leq (1 - \beta_k) \|\bar{\delta}_t^{k-1}\| + \beta_k \mathbb{E}[\|\nabla F_t(x_k) - \nabla F_t(x_{k-1})\|] \\ &\leq (1 - \beta_k) \|\bar{\delta}_t^{k-1}\| + \beta_k L \mathbb{E}[\|x_k - x_{k-1}\|] \\ &\leq (1 - \beta_k) \|\bar{\delta}_t^{k-1}\| + \beta_k L G \eta_{k-1} \end{aligned} \quad (28)$$

Solving the Recurrence Use parameter schedules:

$$\eta_k = \frac{1}{c_\eta k^{2/3}}, \quad \beta_k = \frac{1}{c_\beta k^{2/3}}, \quad \text{let } C := \frac{LG}{c_\eta c_\beta} \quad (29)$$

Then the recurrence becomes:

$$\|\bar{\delta}_t^k\| \leq \left(1 - \frac{1}{c_\beta k^{2/3}}\right) \|\bar{\delta}_t^{k-1}\| + \frac{C}{k^{4/3}} \quad (30)$$

We prove by induction that:

$$\|\bar{\delta}_t^k\| \leq \frac{C_1}{k^{1/3}} \quad (31)$$

Base case: Choose C_1 large enough to cover $\|\bar{\delta}_t^1\| \leq C_1$.

Induction step: Assume $\|\bar{\delta}_t^{k-1}\| \leq \frac{C_1}{(k-1)^{1/3}}$, then

$$\|\bar{\delta}_t^k\| \leq \left(1 - \frac{1}{c_\beta k^{2/3}}\right) \frac{C_1}{(k-1)^{1/3}} + \frac{C}{k^{4/3}}. \quad (32)$$

We want:

$$\left(1 - \frac{1}{c_\beta k^{2/3}}\right) \frac{C_1}{(k-1)^{1/3}} + \frac{C}{k^{4/3}} \leq \frac{C_1}{k^{1/3}}. \quad (33)$$

Using the mean value theorem:

$$\frac{1}{(k-1)^{1/3}} - \frac{1}{k^{1/3}} \leq \frac{1}{3k^{4/3}}. \quad (34)$$

Then:

$$\frac{C_1}{(k-1)^{1/3}} \left(1 - \frac{1}{c_\beta k^{2/3}}\right) \leq \frac{C_1}{k^{1/3}} - \left(\frac{C_1}{c_\beta k^{1/3+2/3}} + \frac{C_1}{3k^{4/3}}\right). \quad (35)$$

So the total is:

$$\|\bar{\delta}_t^k\| \leq \frac{C_1}{k^{1/3}} - \left(\frac{C_1}{3k^{4/3}} + \frac{C_1}{c_\beta k^{5/3}} \right) + \frac{C}{k^{4/3}}. \quad (36)$$

For large enough C_1 , this holds. Hence,

$$\|\bar{\delta}_t^k\| = \mathcal{O}(k^{-1/3}). \quad (37)$$

Lemma 2 (Controlled Gradient Error) *The variance of the stabilized gradient estimator m_t^k is bounded as follows:*

$$\mathbb{E} \left[\|m_t^k - \nabla F_t(x_k) - \bar{\delta}_t^k\|^2 \right] \leq \frac{C_2}{k^{2/3}}. \quad (38)$$

Proof of Lemma 2.

From Lemma 1, we know that:

$$\mathbb{E}[m_t^k] = \nabla F_t(x_k) + \bar{\delta}_t^k. \quad (39)$$

Hence, the expression on the left-hand side can be rewritten as:

$$\mathbb{E} \left[\|m_t^k - \mathbb{E}[m_t^k]\|^2 \right]. \quad (40)$$

This is precisely the definition of the variance of the random vector m_t^k , which we denote as:

$$V_k := \text{Var}(m_t^k). \quad (41)$$

Thus, the main goal is to prove that $\text{Var}(m_t^k) = \mathcal{O}(k^{-2/3})$.

Assumptions and Notations: We adopt the same assumptions and notations as in Lemma 1, with the additional standard assumption:

Bounded Stochastic Gradient Variance: There exists a constant σ^2 such that for any x and t ,

$$\mathbb{E} \left[\|\nabla f_t(x, \xi) - \nabla F_t(x)\|^2 \right] \leq \sigma^2. \quad (42)$$

Recursive Relation of Variance. Our goal is to derive a recursive relationship between $V_k = \text{Var}(m_t^k)$ and $V_{k-1} = \text{Var}(m_t^{k-1})$.

Starting from the recursive definition of m_t^k , we analyze the deviation:

$$\begin{aligned} m_t^k - \mathbb{E}[m_t^k] &= (1 - \beta_k)(m_t^{k-1} + \Delta_k) + \beta_k g_k \\ &\quad - (1 - \beta_k)(\mathbb{E}[m_t^{k-1}] + \mathbb{E}[\Delta_k]) - \beta_k \mathbb{E}[g_k]. \end{aligned} \quad (43)$$

Define the following notations for simplicity:

$$\begin{aligned} \bullet \quad g_k &:= \nabla f_t(x_k, \xi_k) \\ \bullet \quad \Delta_k &:= g_k - \nabla f_t(x_{k-1}, \xi_k) \end{aligned}$$

Reorganizing, we get:

$$\begin{aligned} m_t^k - \mathbb{E}[m_t^k] &= (1 - \beta_k)(m_t^{k-1} - \mathbb{E}[m_t^{k-1}]) \\ &\quad + [(1 - \beta_k)(\Delta_k - \mathbb{E}[\Delta_k]) + \beta_k(g_k - \mathbb{E}[g_k])]. \end{aligned} \quad (44)$$

Let:

$$\begin{aligned} \bullet \quad A &:= (1 - \beta_k)(m_t^{k-1} - \mathbb{E}[m_t^{k-1}]) \\ \bullet \quad B &:= (1 - \beta_k)(\Delta_k - \mathbb{E}[\Delta_k]) + \beta_k(g_k - \mathbb{E}[g_k]) \end{aligned}$$

Since A depends only on past randomness (up to step $k-1$) and B only on current sample ξ_k , we have $\mathbb{E}[\langle A, B \rangle] = 0$.

Using the identity:

$$\mathbb{E}[\|A + B\|^2] = \mathbb{E}[\|A\|^2] + \mathbb{E}[\|B\|^2] + 2\mathbb{E}[\langle A, B \rangle], \quad (45)$$

we obtain a tighter bound:

$$V_k = \mathbb{E}[\|A\|^2] + \mathbb{E}[\|B\|^2] = (1 - \beta_k)^2 V_{k-1} + \mathbb{E}[\|B\|^2]. \quad (46)$$

Using the inequality $\|X + Y\|^2 \leq 2(\|X\|^2 + \|Y\|^2)$, we bound:

$$\begin{aligned} \mathbb{E}[\|B\|^2] &\leq 2(1 - \beta_k)^2 \mathbb{E}[\|\Delta_k - \mathbb{E}[\Delta_k]\|^2] \\ &\quad + 2\beta_k^2 \mathbb{E}[\|g_k - \mathbb{E}[g_k]\|^2]. \end{aligned} \quad (47)$$

Bounding the first term:

$$\begin{aligned} \mathbb{E}[\|\Delta_k - \mathbb{E}[\Delta_k]\|^2] &= \text{Var}(\Delta_k) \leq \mathbb{E}[\|\Delta_k\|^2] \\ &\leq \mathbb{E}[\|\nabla f_t(x_k, \xi_k) - \nabla f_t(x_{k-1}, \xi_k)\|^2] \\ &\leq L^2 \mathbb{E}[\|x_k - x_{k-1}\|^2] \leq L^2 G^2 \eta_k^2. \end{aligned} \quad (48)$$

Bounding the second term:

$$\mathbb{E}[\|g_k - \mathbb{E}[g_k]\|^2] = \text{Var}(g_k) \leq \sigma^2. \quad (49)$$

Plugging into the recursive relation:

$$V_k \leq (1 - \beta_k)^2 V_{k-1} + 2L^2 G^2 \eta_k^2 + 2\beta_k^2 \sigma^2. \quad (50)$$

Since $(1 - \beta_k)^2 \leq 1$, we approximate:

$$(1 - \beta_k)^2 \approx 1 - 2\beta_k, \quad (51)$$

yielding the final recurrence:

$$V_k \leq (1 - 2\beta_k) V_{k-1} + 2L^2 G^2 \eta_k^2 + 2\beta_k^2 \sigma^2. \quad (52)$$

Solving the Recursive Inequality. Assume the following parameter schedules:

- $\beta_k = \frac{1}{c_\beta k^{2/3}}$
- $\eta_k = \frac{1}{c_\eta k^{2/3}}$

Then:

$$V_k \leq \left(1 - \frac{1}{c_\beta k^{2/3}}\right) V_{k-1} + \frac{C'}{k^{4/3}}, \quad (53)$$

where C' is a constant that aggregates the two noise terms.

We aim to prove via induction that:

$$V_k \leq \frac{C_2}{k^{2/3}}, \quad (54)$$

for some constant $C_2 > 0$.

Base Case: For $k = 1$, choose C_2 large enough so that the inequality holds.

Inductive Step: Assume $V_{k-1} \leq \frac{C_2}{(k-1)^{2/3}}$. Then:

$$V_k \leq \left(1 - \frac{1}{c_\beta k^{2/3}}\right) \cdot \frac{C_2}{(k-1)^{2/3}} + \frac{C'}{k^{4/3}}. \quad (55)$$

Using the mean value theorem on $f(x) = x^{-2/3}$, we get:

$$\begin{aligned} (k-1)^{-2/3} - k^{-2/3} &\geq \frac{2}{3k^{5/3}} \\ \Rightarrow k^{-2/3} - (k-1)^{-2/3} &\leq -\frac{2}{3k^{5/3}}. \end{aligned} \quad (56)$$

Thus:

$$V_k \leq \frac{C_2}{k^{2/3}} - \frac{C_2}{c_\beta k^{4/3}} + \frac{C'}{k^{4/3}} = \frac{C_2}{k^{2/3}} + \left(\frac{C' - C_2/c_\beta}{k^{4/3}}\right). \quad (57)$$

For sufficiently large k , the additional term becomes negligible, and the inequality holds if:

$$C_2 \geq c_\beta C'. \quad (58)$$

Hence, by induction:

$$V_k = \mathcal{O}(k^{-2/3}). \quad (59)$$

Theorem 1 Under Lemmas 1 and 2, let the step size be $\eta_k = \eta_0 k^{-1/3}$ for some constant $\eta_0 > 0$. Then there exists a constant $C > 0$ such that:

$$\min_{1 \leq k \leq K} \mathbb{E} \|\nabla F(\mathbf{x}_k)\|^2 \leq \frac{C}{K^{1/3}}, \quad (60)$$

Proof of Theorem 1.

From the L -smoothness of F , we have:

$$\begin{aligned} \mathbb{E}[F(x_{k+1})] &\leq \\ \mathbb{E} \left[F(x_k) + \langle \nabla F(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|^2 \right] \\ &= \mathbb{E} \left[F(x_k) - \eta_k \langle \nabla F(x_k), d_k \rangle + \frac{L\eta_k^2}{2} \|d_k\|^2 \right]. \end{aligned} \quad (61)$$

Rewriting Assumption 1 gives:

$$\begin{aligned} \eta_k \mathbb{E}[\langle \nabla F(x_k), d_k \rangle] &\leq \mathbb{E}[F(x_k)] - \mathbb{E}[F(x_{k+1})] \\ &\quad + \frac{L\eta_k^2}{2} \mathbb{E}[\|d_k\|^2]. \end{aligned} \quad (62)$$

We expand the inner product:

$$\begin{aligned} \mathbb{E}[\langle \nabla F(x_k), d_k \rangle] &= \mathbb{E}[\|\nabla F(x_k)\|^2] \\ &\quad + \mathbb{E}[\langle \nabla F(x_k), d_k - \nabla F(x_k) \rangle]. \end{aligned} \quad (63)$$

By Young's inequality:

$$\begin{aligned} \mathbb{E}[\langle \nabla F(x_k), d_k - \nabla F(x_k) \rangle] &\geq -\frac{1}{2} \mathbb{E}[\|\nabla F(x_k)\|^2] \\ &\quad - \frac{1}{2} \mathbb{E}[\|d_k - \nabla F(x_k)\|^2]. \end{aligned} \quad (64)$$

Thus,

$$\begin{aligned} \mathbb{E}[\langle \nabla F(x_k), d_k \rangle] &\geq \frac{1}{2} \mathbb{E}[\|\nabla F(x_k)\|^2] \\ &\quad - \frac{1}{2} \mathbb{E}[\|d_k - \nabla F(x_k)\|^2]. \end{aligned} \quad (65)$$

Now decompose the squared error term:

$$\begin{aligned} \mathbb{E}[\|d_k - \nabla F(x_k)\|^2] &= \text{Var}(d_k) + \|\mathbb{E}[d_k] - \nabla F(x_k)\|^2 \\ &\leq \frac{C_2}{k^{2/3}} + \frac{C_1^2}{k^{2/3}} = \frac{C_3}{k^{2/3}}. \end{aligned} \quad (66)$$

Assume $\mathbb{E}[\|d_k\|^2] \leq G_d^2$. Plugging (65) and (66) into (62):

$$\begin{aligned} \frac{\eta_k}{2} \mathbb{E}[\|\nabla F(x_k)\|^2] &\leq \mathbb{E}[F(x_k)] - \mathbb{E}[F(x_{k+1})] \\ &\quad + \frac{\eta_k C_3}{2k^{2/3}} + \frac{L\eta_k^2 G_d^2}{2}. \end{aligned} \quad (67)$$

Using $\eta_k = \eta_0 k^{-1/3}$, we get:

$$\begin{aligned} \frac{\eta_0}{2k^{1/3}} \mathbb{E}[\|\nabla F(x_k)\|^2] &\leq \mathbb{E}[F(x_k)] - \mathbb{E}[F(x_{k+1})] \\ &\quad + \frac{\eta_0 C_3}{2k} + \frac{L\eta_0^2 G_d^2}{2k^{2/3}}. \end{aligned} \quad (68)$$

Summing from $k = 1$ to K :

$$\begin{aligned} \sum_{k=1}^K \frac{\eta_0}{2k^{1/3}} \mathbb{E}[\|\nabla F(x_k)\|^2] &\leq F(x_1) - F^* \\ &\quad + \sum_{k=1}^K \frac{\eta_0 C_3}{2k} + \sum_{k=1}^K \frac{L\eta_0^2 G_d^2}{2k^{2/3}}. \end{aligned} \quad (69)$$

Method	Segmentation \uparrow		Depth \downarrow		$\Delta m\%$ \downarrow	Cost
	mIoU	Pix. Acc.	Abs. Err.	Rel. Err.		
FairGrad	75.72	93.68	0.0134	32.25	5.18	100%
VarGrad (SMO)	74.75	93.45	0.0145	30.37	5.90	44.84%

Table 5. MTL update efficiency on *Cityscapes* (2 tasks) dataset.

Note:

- $\sum_{k=1}^K \frac{1}{k} = \mathcal{O}(\log K)$
- $\sum_{k=1}^K \frac{1}{k^{2/3}} = \mathcal{O}(K^{1/3})$
- $\sum_{k=1}^K \frac{1}{k^{1/3}} = \mathcal{O}(K^{2/3})$

Hence:

$$\left(\min_{1 \leq k \leq K} \mathbb{E}[\|\nabla F(x_k)\|^2] \right) \cdot \sum_{k=1}^K \frac{\eta_0}{2k^{1/3}} \leq \mathcal{O}(K^{1/3}), \quad (70)$$

which implies:

$$\min_{1 \leq k \leq K} \mathbb{E}[\|\nabla F(x_k)\|^2] \leq \frac{\mathcal{O}(K^{1/3})}{\mathcal{O}(K^{2/3})} = \frac{C}{K^{1/3}}. \quad (71)$$

7.3. C. Additional Results

7.3.1. Efficiency Analysis.

This experiment evaluates the efficiency of multi-task updates using VarGrad on the *Cityscapes* dataset with two tasks: semantic segmentation and depth estimation. As shown in Table 4, VarGrad achieves competitive task performance compared to the baseline FairGrad—slightly lower mIoU (74.75 vs. 75.72) and comparable depth error—with only minor differences across metrics. Crucially, VarGrad performs multi-task coordination (i.e., MTL-specific updates) in only **44.84%** of the training steps, whereas FairGrad applies it at every step (100%). This reflects the effectiveness of our selective update strategy (SMO): by adaptively skipping coordination when task progress is balanced, VarGrad reduces computational overhead without sacrificing performance.

Overall, these results demonstrate that VarGrad enables more efficient training by focusing coordination efforts where they are most needed, offering a practical advantage in real-world multi-task learning scenarios.

7.3.2. Hyper-parameter Sensitivity Analysis.

These results present an ablation study on the sensitivity of VarGrad to the hyperparameter β , which controls the strength of variance reduction. As shown in the table, performance varies with different β values. When β is set too low (e.g., 0.75), the effect of variance reduction is weaker, leading to less consistent improvements across tasks ($\Delta m\% = 4.70$). Conversely, setting β too high (e.g., 0.95) results in over-suppression of gradient signals, which harms

Beta	Segmentation \uparrow		Depth \downarrow		$\Delta m\%$ \downarrow
	mIoU	Pix. Acc.	Abs. Err.	Rel. Err.	
Beta=0.75	74.68	93.39	0.0134	31.38	4.70
Beta=0.8	75.20	93.38	0.0146	29.92	5.72
Beta=0.85	75.22	93.52	0.0124	31.05	2.33
Beta=0.9	74.50	93.39	0.0138	32.33	6.40
Beta=0.95	75.48	93.48	0.0135	38.18	10.87

Table 6. Impact of Hyper-parameter β on *Cityscapes* (2 tasks) dataset.

depth performance and significantly increases task imbalance ($\Delta m\% = 10.87$).

The best overall performance is achieved with $\beta = 0.85$, yielding the lowest $\Delta m\%$ (2.33) and strong task metrics, indicating a good balance between smoothing and preserving informative gradients. This validates the importance of tuning β to strike an effective trade-off.

7.4. Limitations

While our method shows consistent improvements across diverse MTL benchmarks, it is not without limitations:

7.4.1. Task- and Method-Dependent Gains.

The effectiveness of our approach can vary across different datasets and MTL baselines. For instance, when gradient variance is already low (e.g., due to large batch sizes or highly correlated tasks), the benefits of variance reduction become less pronounced. Similarly, the degree of improvement may depend on how well the underlying MTL optimizer (e.g., MGDA, FairGrad) interacts with our variance correction.

7.4.2. Manual Hyperparameter Selection.

Our method introduces task-agnostic hyperparameters such as the variance smoothing coefficient β and the coordination threshold τ , which we select via grid search. However, their optimal values can vary across datasets or task types. This limits adaptivity and may require manual tuning for new applications. A more principled or dynamic hyperparameter selection mechanism could improve usability.