

Supplementary Materials: AxiomVision: Accuracy-Guaranteed Adaptive Visual Model Selection for Perspective-Aware Video Analytics

Xiangxiang Dai
The Chinese University of Hong Kong, Hong Kong, China
xxdai23@cse.cuhk.edu.hk

Zeyu Zhang*
Huazhong University of Science and Technology, Wuhan, China
zeyuzhangzyz@gmail.com

Peng Yang
Huazhong University of Science and Technology, Wuhan, China
yangpeng@hust.edu.cn

Yuedong Xu
Fudan University, Shanghai, China
ydxu@fudan.edu.cn

Xutong Liu†
The Chinese University of Hong Kong, Hong Kong, China
liuxt@cse.cuhk.edu.hk

John C.S. Lui
The Chinese University of Hong Kong, Hong Kong, China
cslui@cse.cuhk.edu.hk

Appendix

A Proof of Theorem 1

PROOF. For any processed camera $n \in \mathcal{N}$, we define the Gramian matrix and the number of effective feedbacks for camera n up to round t , respectively, as follows:

$$\mathbf{M}_{n,t} = \sum_{\substack{j \leq t \\ n_j = n}} \sum_{k=1}^{|\mathcal{K}_j|} \mathbf{x}_{m_k,j} \mathbf{x}_{m_k,j}^\top, \quad T_{n,t} = \sum_{\substack{j \leq t \\ n_j = n}} |\mathcal{K}_j|$$

Subsequently, for any camera n belonging to group index i , denote

$$\mathbf{M}_{i,t} = \zeta \mathbf{I}_d + \sum_{n \in G_i} \mathbf{M}_{n,t}, \quad T_{i,t} = \sum_{n \in G_i} T_{n,t} \quad (1)$$

be the regularized Gramian matrix, and the frequency associated with belonging group G_i , respectively, incorporating the regularization parameter $\zeta > 0$ up to round t .

Consider the gradient function defined for any camera within group G_i at time t as:

$$g_{i,t}(\boldsymbol{\theta}) = \sum_{j=1}^{t-1} \mathbf{1}\{n_j \in G_i\} \sum_{k=1}^{|\mathcal{K}_j|} \mu(\mathbf{x}_{m_k,j}^\top \boldsymbol{\theta}) \mathbf{x}_{m_k,j}.$$

Recall that $\hat{\boldsymbol{\theta}}_{i,t}$, t is identified as the unique solution of the equation:

$$\sum_{j=1}^{t-1} \mathbf{1}\{n_j \in G_i\} \sum_{k=1}^{|\mathcal{K}_j|} \left(r_{m_k,j} - \mu(\mathbf{x}_{m_k,j}^\top \hat{\boldsymbol{\theta}}_{i,t}) \right) \mathbf{x}_{m_k,j} = 0.$$

Then, it is possible to express $g_{i,t}(\hat{\boldsymbol{\theta}}_{i,t-1})$, which captures the cumulative response adjusted by the previous estimate of $\boldsymbol{\theta}$, as follows: $g_{i,t}(\hat{\boldsymbol{\theta}}_{i,t-1}) = \sum_{s=1}^{t-1} \mathbf{1}\{n_s \in G_i\} \sum_{k=1}^{|\mathcal{K}_s|} r_{m_k,s} \mathbf{x}_{m_k,s}$. This formulation integrates the feedback up to round $t-1$, weighted by the membership of the cameras in the group G_i , to refine the estimation of the parameter $\boldsymbol{\theta}$.

Next, we introduce a lemma that provides a theoretical guarantee for the accuracy of the ridge regression estimate in approximating the true weight vector of camera perspective influence. This lemma is critical for understanding the bounds of estimation error in linear models with ridge regression.

LEMMA 1 (THEOREM 1 IN [5]). *Consider a sequence of data points $(x_1, y_1), \dots, (x_t, y_t)$ are generated sequentially from a linear model such that $\|x_t\| \leq 1$ for all t , $\mathbb{E}[y_t | x_t] = \theta_*^\top x_t$ for fixed but unknown θ_* with norm at most 1, and $\{y_t - \theta_*^\top x_t\}_{t=1,2,\dots}$ have 1-sub-Gaussian tails. Let $M_t = \zeta \mathbf{I} + \sum_{s=1}^t x_s x_s^\top$, $\mathbf{b}_t = \sum_{s=1}^t x_s y_s$, and $\delta > 0$. If $\hat{\boldsymbol{\theta}}_t = M_t^{-1} \mathbf{b}_t$ is the ridge regression estimator of θ_* , then with probability at least $1 - \delta$, for all $t \geq 0$,*

$$\|\hat{\boldsymbol{\theta}}_t - \theta_*\|_{M_t} \leq \sqrt{d \ln \left(1 + \frac{t}{\zeta d} \right)} + 2 \ln \frac{1}{\delta} + \sqrt{\zeta}.$$

Furthermore, we examine the determinant of the matrix M_t and its relationship with its eigenvalues and trace, leading to:

$$\det(M_t) \leq \left(\frac{\sum_{i=1}^d \lambda_i}{d} \right)^d = \left(\frac{\text{trace}(M_t)}{d} \right)^d \leq \left(\frac{t + \zeta d}{d} \right)^d,$$

where λ_i , ($i = 1, 2, \dots, d$) denotes the eigenvalues of the matrix M_t , $\text{trace}(M_t)$ denotes the trace of M_t . By Lemma 1, combined with the inequality: $\det(M_{i,t}) \leq \frac{1}{d^d} \left(\text{trace}(M_{i,t}) + \sum_{j=1}^t \|x_j\|_2^2 \right)^d \leq (\zeta + t/d)^d$, for some $j \leq t$ (will be clarified later) with $M_{i,j}$ invertible, with probability at least $1 - \delta$, we have:

$$\|g_{i,t}(\hat{\boldsymbol{\theta}}_{i,t}) - g_{i,t}(\boldsymbol{\theta}_{i,t})\|_{M_{i,t}^{-1}}^2 \leq T_{i,t} \lambda_{\min}(M_{i,t})^{-1} + d \ln \frac{T_{i,t}}{d} + 2 \ln \frac{1}{\delta}.$$

Here, $\lambda_{\min}(M)$ denotes the minimum eigenvalue of matrix M .

Leveraging the Lipschitz continuity and the properties of the first derivative for the function μ , $g_{i,t}(\hat{\boldsymbol{\theta}}_{i,t}) - g_{i,t}(\boldsymbol{\theta}_{i,t}) \geq m_\mu M_{i,t}$, we assert that the difference between the gradient functions evaluated at the estimated and true parameter vectors is bounded by the product of m_μ and the matrix $M_{i,t}$. Consequently, this relationship yields the following inequality:

$$m_\mu^2 \|\hat{\boldsymbol{\theta}}_{i,t} - \boldsymbol{\theta}_{i,t}\|_{M_{i,t}}^2 \leq T_{i,t} \lambda_{\min}(M_{i,t})^{-1} + d \ln \frac{T_{i,t}}{d} + 2 \ln \frac{1}{\delta}.$$

From Lemma 1, denoting $\boldsymbol{\alpha}(t, \delta) = \frac{1}{m_\mu} \sqrt{\frac{8}{\lambda} + d \ln \frac{t}{d} + 2 \ln \frac{1}{\delta}}$, we can derive that the following inequality holds:

$$\begin{aligned} \left| \mu(\mathbf{x}_{m_t}^\top \hat{\boldsymbol{\theta}}_{i,t}) - \mu(\mathbf{x}_{m_t}^\top \boldsymbol{\theta}_{i,t}) \right| &\leq L \left\| \mathbf{x}_{m_t}^\top \hat{\boldsymbol{\theta}}_{i,t} - \mathbf{x}_{m_t}^\top \boldsymbol{\theta}_{i,t} \right\| \\ &\leq 2L \boldsymbol{\alpha}(T_{i,t-1}, \delta) \|\mathbf{x}_{m_t}\|_{M_{i,t}^{-1}}. \end{aligned}$$

*Work conducted during Zeyu Zhang's visit to The Chinese University of Hong Kong.
†Xutong Liu is the corresponding author.

which confirms the accuracy of the estimated group index i_t for camera n_t . Here, m_t^* represents the optimal model at round t .

In the context of this analysis, let us consider a scenario where for any $\omega > 0$ and any $m \in \mathcal{M}_t$:

$$\mathbb{P}_t(\boldsymbol{\theta}_m < -\omega | \mathcal{M}_t) \leq e^{-\frac{\omega^2}{2\sigma^2}},$$

where $\mathbb{P}_t(\cdot)$ is the shorthand for the conditional probability given the size of the set \mathcal{M}_t . Given that

$$\begin{aligned} \mathbb{E}_t[(\boldsymbol{\theta}^\top \mathbf{x}_{m,t})^2 | \mathcal{M}_t] &= \mathbb{E}_t[\boldsymbol{\theta}^\top \mathbf{x}_{m,t} \mathbf{x}_{m,t}^\top \boldsymbol{\theta} | \mathcal{M}_t] \\ &\geq \lambda_{\min}(\mathbb{E}_{\mathbf{x} \sim \rho}[\mathbf{x}\mathbf{x}^\top]) \geq \lambda, \end{aligned}$$

it follows that

$$\mathbb{P}_t\left(\min_{i=1, \dots, |\mathcal{M}_t|} (\boldsymbol{\theta}^\top \mathbf{x}_{m,t})^2 \geq \lambda - \omega | \mathcal{M}_t\right) \geq (1 - e^{-\frac{\omega^2}{2\sigma^2}})^K.$$

From this, we deduce that

$$\begin{aligned} \mathbb{E}_t[(\boldsymbol{\theta}^\top \mathbf{x}_{m,t})^2 | \mathcal{M}_t] &\geq \mathbb{E}_t\left[\min_{m \in \mathcal{M}_t} (\boldsymbol{\theta}^\top \mathbf{x}_{m,t})^2 | \mathcal{M}_t\right] \\ &\geq \int_0^\infty \mathbb{P}_t\left(\min_{m \in \mathcal{M}_t} (\boldsymbol{\theta}^\top \mathbf{x}_{m,t})^2 \geq x | \mathcal{M}_t\right) dx \\ &\geq \int_0^\lambda (1 - e^{-\frac{(\lambda-x)^2}{2\sigma^2}})^K dx \triangleq \tilde{\lambda} \end{aligned}$$

This establishes a lower bound, $\tilde{\lambda}$, on the expected squared projection of $\boldsymbol{\theta}$ onto the feature vectors $\mathbf{x}_{m,t}$ for any m in \mathcal{M}_t ,

By Claim 1 of [17], Lemma 7, 8 of [39], for each camera n , for all $T_{n,t} \geq \frac{1024}{\lambda^2} \ln \frac{512d}{\lambda^2 \delta}$, $\lambda_{\min}(\mathbf{M}_{n,t}) \geq T_{n,t} \tilde{\lambda} / 8$ with high probability. Thus with high probability, for the belonging group index i , we have:

$$\begin{aligned} \|\hat{\boldsymbol{\theta}}_{n,t} - \boldsymbol{\theta}_n\|_{\mathbf{M}_{i,t}} &\leq \frac{1}{m_\mu} \sqrt{\frac{8}{\tilde{\lambda}} + d \ln \frac{T_{n,t}}{d} + 2 \ln \frac{1}{\delta}} \\ \|\hat{\boldsymbol{\theta}}_{n,t} - \boldsymbol{\theta}_n\| &\leq \frac{\sqrt{\frac{8}{\tilde{\lambda}} + d \ln \frac{T_{n,t}}{d} + 2 \ln \frac{1}{\delta}}}{m_\mu \sqrt{\tilde{\lambda} T_{n,t} / 8}}. \end{aligned}$$

When $T_{n,t} \geq \frac{512d}{(\gamma q)^2 \lambda} \ln \frac{|\mathcal{N}|}{\delta}$ with Lemma 10 in [39], we have: $\frac{\alpha(T_{n,t}, \delta)}{m_\mu \sqrt{\tilde{\lambda} T_{n,t} / 8}}$

$< \frac{\gamma q}{4}$. According to Lemma 1 in [66], these conditions are satisfied with a probability of at least $1 - \delta$, for δ in the interval $(0, 1)$, provided that the time t meets or exceeds

$$\begin{aligned} t \geq &4|\mathcal{N}| \max \left\{ \frac{512d}{(\gamma q)^2 \lambda} \ln \frac{|\mathcal{N}|}{\delta}, \frac{256}{\lambda^2} \ln \frac{32d}{\lambda^2 \delta} \right\} \\ &+ 16|\mathcal{N}| \ln \frac{4|\mathcal{N}|T}{\delta} =: T_{q,0}. \end{aligned} \quad (2)$$

With this, we can show that *AxiomVision* will group all the cameras correctly after $T_{q,0}$. Let $\alpha = \frac{16\sqrt{d}}{\lambda m_\mu}$ and use $i(n)$ to denote the index of group camera n belongs to (i.e., $n \in G_{i(n)}$). First, if *AxiomVision* deletes the edge (n, ℓ) , then camera n and camera ℓ belong to different ground-truth groups, i.e., $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_\ell\|_2 > 0$. This is because by the deletion rule of the algorithm, the concentration bound, and triangle inequality, $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_\ell\|_2 \geq \|\hat{\boldsymbol{\theta}}_{n,t} - \hat{\boldsymbol{\theta}}_{\ell,t}\|_2 - \|\boldsymbol{\theta}_{i(n)} - \boldsymbol{\theta}_{i(\ell)}\|_2 -$

$\|\boldsymbol{\theta}_{i(n)} - \boldsymbol{\theta}_{n,t}\|_2 > 0$. Second, we show that if $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_\ell\| \geq \gamma q$, *AxiomVision* will delete the edge (n, ℓ) , where the dispersion condition in Eq. (3) implies that if the condition of deleting edge in Eq. (7) is met, then cameras n, ℓ indeed belong to different ground-truth groups. This is because if $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_\ell\| \geq \gamma q$, then by the triangle inequality, and $\|\hat{\boldsymbol{\theta}}_{n,t} - \boldsymbol{\theta}_{i(n)}\|_2 < \frac{\gamma q}{4}$, $\|\hat{\boldsymbol{\theta}}_{\ell,t} - \boldsymbol{\theta}_{i(\ell)}\|_2 < \frac{\gamma q}{4}$, $\boldsymbol{\theta}_n = \boldsymbol{\theta}_{i(n)}$, $\boldsymbol{\theta}_\ell = \boldsymbol{\theta}_{i(\ell)}$, we have $\|\hat{\boldsymbol{\theta}}_{n,t} - \hat{\boldsymbol{\theta}}_{\ell,t}\|_2 \geq \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_\ell\| - \|\hat{\boldsymbol{\theta}}_{n,t} - \boldsymbol{\theta}_{i(n)}\|_2 - \|\hat{\boldsymbol{\theta}}_{\ell,t} - \boldsymbol{\theta}_{i(\ell)}\|_2 > \gamma q - \frac{\gamma q}{4} - \frac{\gamma q}{4} = \frac{\gamma q}{2}$. Note that the threshold $\beta(f(T_{n,t,t-1}) + f(T_{\ell,t,t-1}))$ in Eq. (7) is designed in accordance with the theoretical framework [17]. Moreover, Algorithm 1 reinstates the fully connected graph visual model as detailed in Line 11. This adjustment ensures that, when applied across all cameras, the derived groupings are precise.

Once *AxiomVision* successfully groups all cameras accurately, it follows that the discrepancy between the estimated parameter vector $\hat{\boldsymbol{\theta}}_{i_t,t-1}$ for group i_t at time $t-1$ and the true parameter vector $\boldsymbol{\theta}_{n_t}$ for camera n_t is bounded by the norm induced by the matrix $\mathbf{M}_{i_t,t-1}$:

$$\|\hat{\boldsymbol{\theta}}_{i_t,t-1} - \boldsymbol{\theta}_{n_t}\|_{\mathbf{M}_{i_t,t-1}} \leq \alpha(T_{i_t,t-1}, \delta) \leq \alpha(T, \delta). \quad (3)$$

Then, applying the Cauchy-Schwarz inequality, we obtain:

$$\begin{aligned} &|\mathbf{x}_{m_t}^\top (\hat{\boldsymbol{\theta}}_{i_t,t-1} - \boldsymbol{\theta}_{n_t})| \\ &\leq \|\mathbf{x}_{m_t}\|_{\mathbf{M}_{i_t,t-1}^{-1}} \|\hat{\boldsymbol{\theta}}_{i_t,t-1} - \boldsymbol{\theta}_{n_t}\|_{\mathbf{M}_{i_t,t-1}} \leq \alpha(T, \delta) \|\mathbf{x}_{m_t}\|_{\mathbf{M}_{i_t,t-1}^{-1}}. \end{aligned}$$

By setting $\delta = 1/T$, with probability at least $1 - 1/T$, for each camera n_t with the selected visual model m_t under query q , the following holds:

$$|\mathbf{x}_{m_t}^\top (\hat{\boldsymbol{\theta}}_{i_t,t-1} - \boldsymbol{\theta}_{n_t})| \leq \alpha \|\mathbf{x}_{m_t}\|_{\mathbf{M}_{i_t,t-1}^{-1}}. \quad (4)$$

This establishes a confidence interval that bridges the gap between group-based visual model selection and individual selection.

The analysis of instantaneous regret, Reg_t , at any given round t for the selected visual model m_t under visual task q , allows us to assess the immediate loss due to the choice of model at that round. For $t \geq T_{q,0}$, the instantaneous regret is given by:

$$\begin{aligned} Reg_t &= \mu(\mathbf{x}_{m_t}^\top \boldsymbol{\theta}_{n_t}) - \mu(\mathbf{x}_{m_t}^\top \boldsymbol{\theta}_{n_t}) \\ &\leq 2\alpha L(T_{i_t,t-1}, \delta) \|\mathbf{x}_{m_t}\|_{\mathbf{M}_{i_t,t-1}^{-1}} \leq 2\alpha L \|\mathbf{x}_{m_t}\|_{\mathbf{M}_{i_t,t-1}^{-1}}. \end{aligned} \quad (5)$$

Aggregating all instant regrets under visual task q into a total regret $Reg(T_q)$, and applying the Cauchy-Schwarz inequality for

summation over all rounds, we derive:

$$\begin{aligned}
 \text{Reg}(T_q) &= \mathbb{E} \left[\sum_{t=1}^{T_{q,0}} \mathbb{E}_t (\text{Reg}_t) \right] + \mathbb{E} \left[\sum_{t=T_{q,0}+1}^{T_q} \mathbb{E}_t (\text{Reg}_t) \right] \\
 &\leq T_{q,0} + \sum_{t=T_{q,0}+1}^{T_q} L \min \left\{ 2, 2\alpha \|x_{m_t}\|_{M_{t,t-1}^{-1}} \right\} \\
 &\stackrel{(a)}{\leq} T_{q,0} + 2\alpha \sum_1^{g_q} \sqrt{T_q \sum_{t=T_{q,0}+1}^{T_q} \min\{1, \|x_{m_t}\|_{M_{t,t-1}^{-1}}^2\}} \\
 &\stackrel{(b)}{\leq} T_{q,0} + 2\alpha L \sqrt{2g_q d T \log \left(1 + \frac{T}{\zeta d} \right)},
 \end{aligned} \tag{6}$$

where g_q denotes the number of ground-truth camera groups under visual task q and (a) is due to the Cauchy-Schwarz inequality, (b) is by Lemma 11 in [5]. Furthermore, in the event of failure, which occurs with a probability of at most δ , the regret remains constant. \square

B Extended Experiments

B.1 Effect of Camera Angles on Model Selection

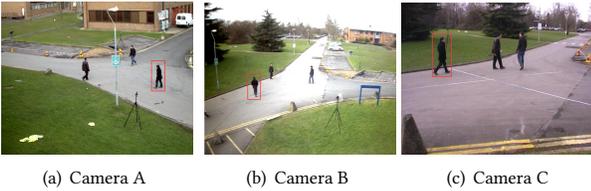


Figure 1: The same object from different perspectives.

Fig. 1 demonstrates how the perception of the same object changes from different camera angles. Camera A, positioned at a greater distance and higher angle, provides an overview perspective, potentially altering the perception of size and shape due to increased distance and angle, thereby necessitating a more complex object detection algorithm. On the other hand, Camera C captures the object at a low angle and close range, clearly revealing details, which allows for the use of a more lightweight object detection algorithm. These variations in perspective significantly impact how surveillance or computer vision systems detect and recognize objects, underscoring the influence of different camera angles on the selection of visual models.

B.2 Accuracy-Bandwidth Trade-off

As illustrated in Fig. 2, our system adeptly achieves a balance between two pivotal metrics—accuracy and bandwidth—through an online visual model selection process based on the tiered edge-cloud architecture. In comparison, both *EAMU* and *Chameleon* limit themselves to relatively uniform visual models which result in either compromised accuracy or excessive bandwidth consumption. Conversely, by setting combinatorial accuracy thresholds of $Th = 0.7, 0.8, 0.9$ under various bandwidth scenarios respectively,

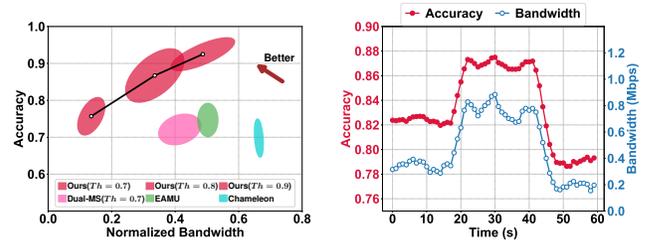


Figure 2: Accuracy v.s. the normalized bandwidth.

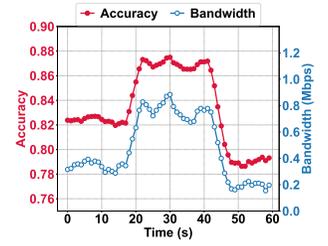


Figure 3: Accuracy under varying bandwidth.

AxiomVision consistently achieves higher accuracy with lower bandwidth usage. This efficiency allows for the conservation of bandwidth resources without sacrificing accuracy.

B.3 Accuracy with Bandwidth Variability

Fig. 3 depicts the behavior of video feeds over a 60-second span, subject to dynamic and varied bandwidth conditions at ENs. Notably, Fig. 3 highlights how *AxiomVision* dynamically adjusts its selection of visual models in response to a surge in bandwidth availability between 20 and 40 seconds. This adjustment enables the adoption of more resource-demanding visual models, resulting in a significant boost in accuracy. Such adaptability demonstrates *AxiomVision*'s exceptional responsiveness to changes in bandwidth. For detailed changes in accuracy for the four specific visual tasks as bandwidth varies, please refer to Table 1-4.

B.4 Ablation Study on Different Parameters

Fig. 4 and Fig. 5 present a detailed comparative analysis on both fixed and adjustable perspective cameras, illustrating the performance of our proposed system, *AxiomVision*, against *Dual-MS* across four distinct visual tasks. This thorough comparison, conducted over a wide range of parameter settings, ensures a comprehensive evaluation of the system's capabilities. Notably, the superior accuracy and consistent performance of *AxiomVision* underscore its robustness and adaptability to various environmental conditions.

B.5 Visual Model Deployment Analysis

We continue to explore deployment strategies for visual models within edge-cloud architectures. Specifically, we categorize the deployed visual models into three levels: simple, medium-complexity, and complex visual models. The simple and medium-complexity models are deployed at the ENs, while the complex models are deployed in the cloud. This comprehensive tiering strategy for visual model deployment is named *AxiomVision (multi-level)*. In the main text, we also introduce a deployment strategy that includes only two levels of models, referred to as *Dual-MS*. Here, we explore the comparison between different dual-level models and multi-level models. Specifically, we have explored the advantages and disadvantages of models resulting from different selection methods.

First, we introduce a dual-level deployment strategy that incorporates only the simple and medium-complexity models, referred to as *AxiomVision (bi-level1)*. As shown in Fig. 6(a)-(c), we initially assessed the impact of these two deployment strategies on accuracy.

Table 1: Accuracy under varying bandwidth for classification.

Time (s)	0	4	8	12	16	20	24	28	32	36	40	44	48	52	56
Accuracy	0.922	0.922	0.924	0.916	0.920	0.918	0.920	0.921	0.919	0.913	0.925	0.930	0.928	0.931	0.933
Bandwidth (Mbps)	0.407	0.365	0.365	0.267	0.232	0.203	0.147	0.161	0.168	0.098	0.168	0.182	0.175	0.196	0.224

Table 2: Accuracy under varying bandwidth for counting.

Time (s)	0	4	8	12	16	20	24	28	32	36	40	44	48	52	56
Accuracy	0.923	0.923	0.929	0.921	0.920	0.926	0.960	0.962	0.965	0.969	0.967	0.928	0.898	0.902	0.890
Bandwidth (Mbps)	1.024	1.270	1.228	1.094	1.144	1.144	1.508	1.768	1.789	1.726	1.628	1.186	0.779	0.954	0.730

Table 3: Accuracy under varying bandwidth for detection.

Time (s)	0	4	8	12	16	20	24	28	32	36	40	44	48	52	56
Accuracy	0.913	0.912	0.910	0.906	0.907	0.942	0.979	0.972	0.978	0.976	0.880	0.835	0.845	0.830	0.843
Bandwidth (Mbps)	1.515	1.593	1.501	1.431	1.557	1.985	2.540	2.392	2.526	2.441	1.256	0.702	0.898	0.695	0.716

Table 4: Accuracy under varying bandwidth for aggregation.

Time (s)	0	4	8	12	16	20	24	28	32	36	40	44	48	52	56
Accuracy	0.824	0.823	0.827	0.823	0.822	0.855	0.870	0.871	0.869	0.865	0.871	0.835	0.790	0.786	0.792
Bandwidth (Mbps)	0.316	0.351	0.379	0.316	0.358	0.631	0.772	0.821	0.751	0.673	0.765	0.519	0.168	0.210	0.210

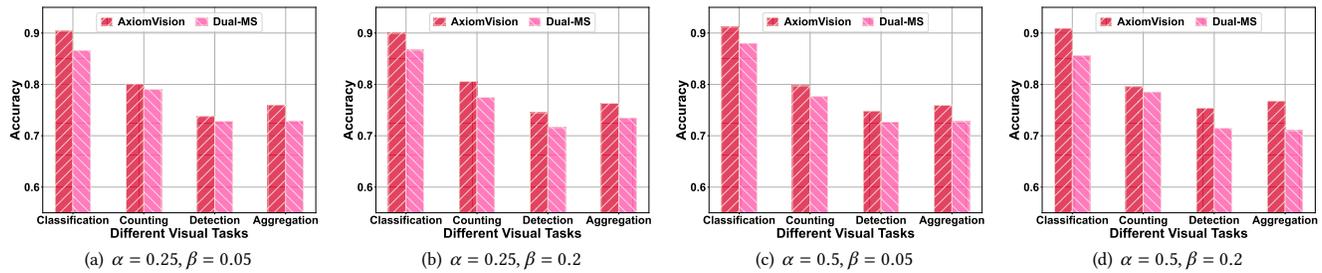


Figure 4: Comparative analysis of AxiomVision performance on fixed perspectives with different parameters.

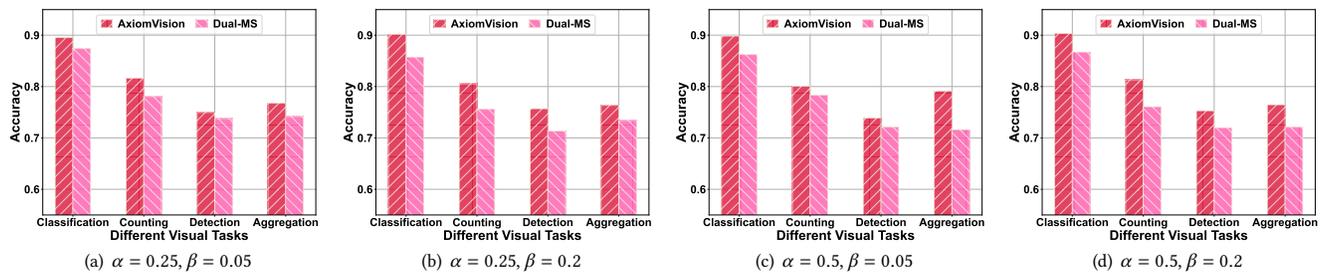


Figure 5: Comparative analysis of AxiomVision performance on adjustable perspectives with different parameters.

The results indicate that *AxiomVision (multi-level)* outperforms *AxiomVision (bi-level1)* across various object detection tasks, demonstrating higher accuracy. Additionally, we considered bandwidth factors, taking into account a trade-off between accuracy and cost, denoted as $a - \eta b$, where a and b represent normalized accuracy

and bandwidth, respectively, and $\eta = 0.5$ is a weighted unit conversion parameter. Fig. 6(d)-(f) reveals that although *AxiomVision (bi-level1)*, by not requiring the scheduling of complex cloud-based visual models, saves bandwidth resources, its inability to handle

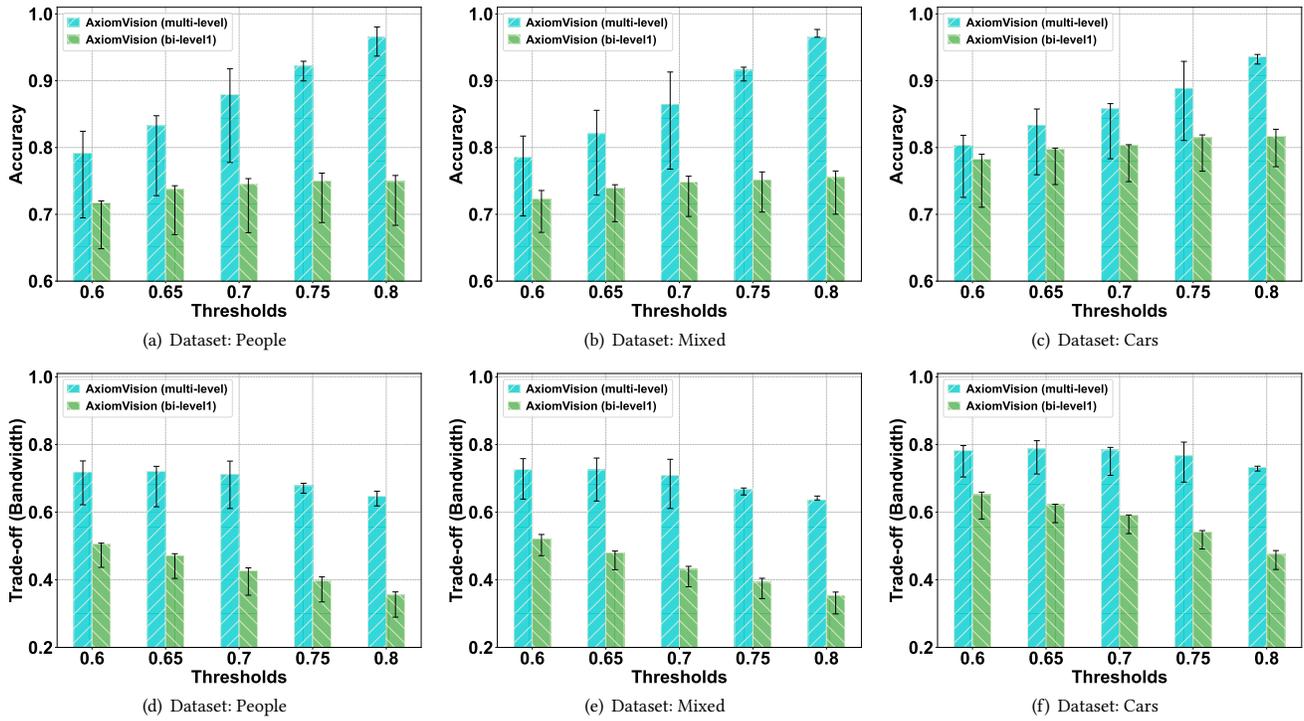


Figure 6: Comparison of *AxiomVision* with multi deployment levels; *AxiomVision* (bi-level) with the first and second levels.

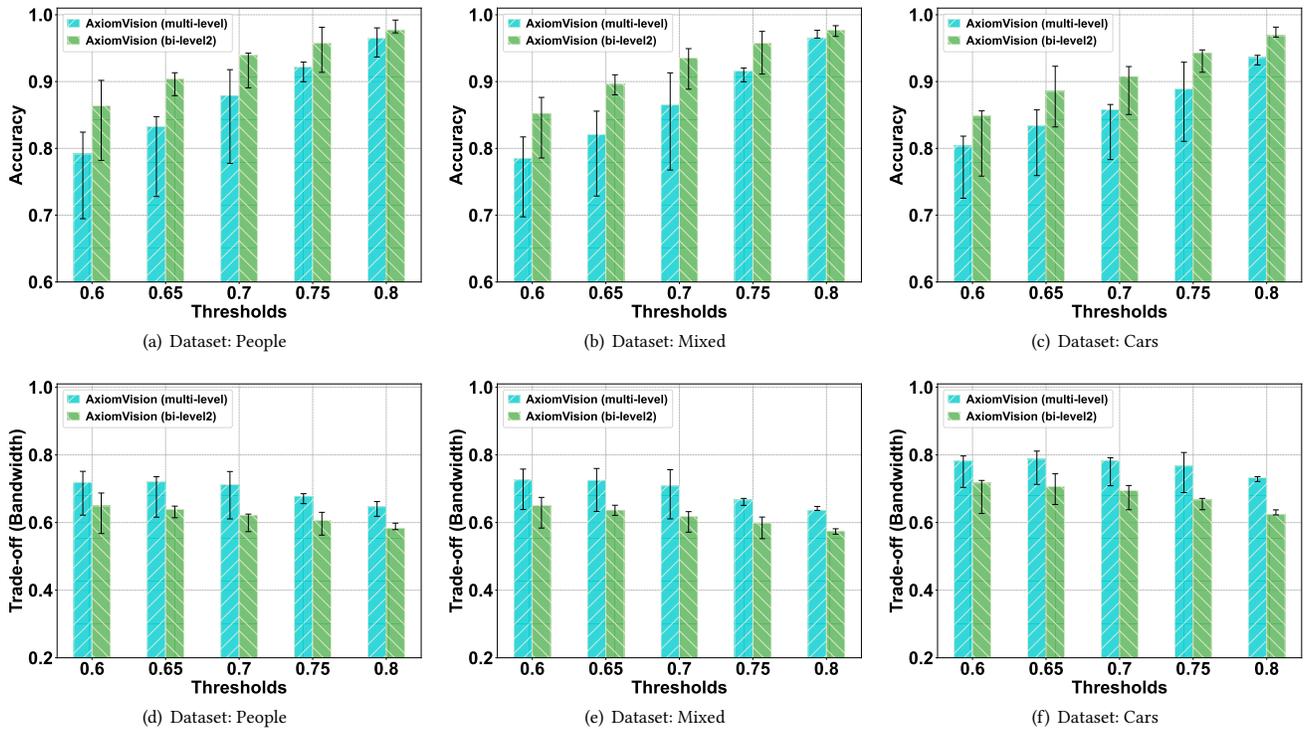


Figure 7: Comparison of *AxiomVision* with multi deployment levels; *AxiomVision* (bi-level) with the first and third levels.

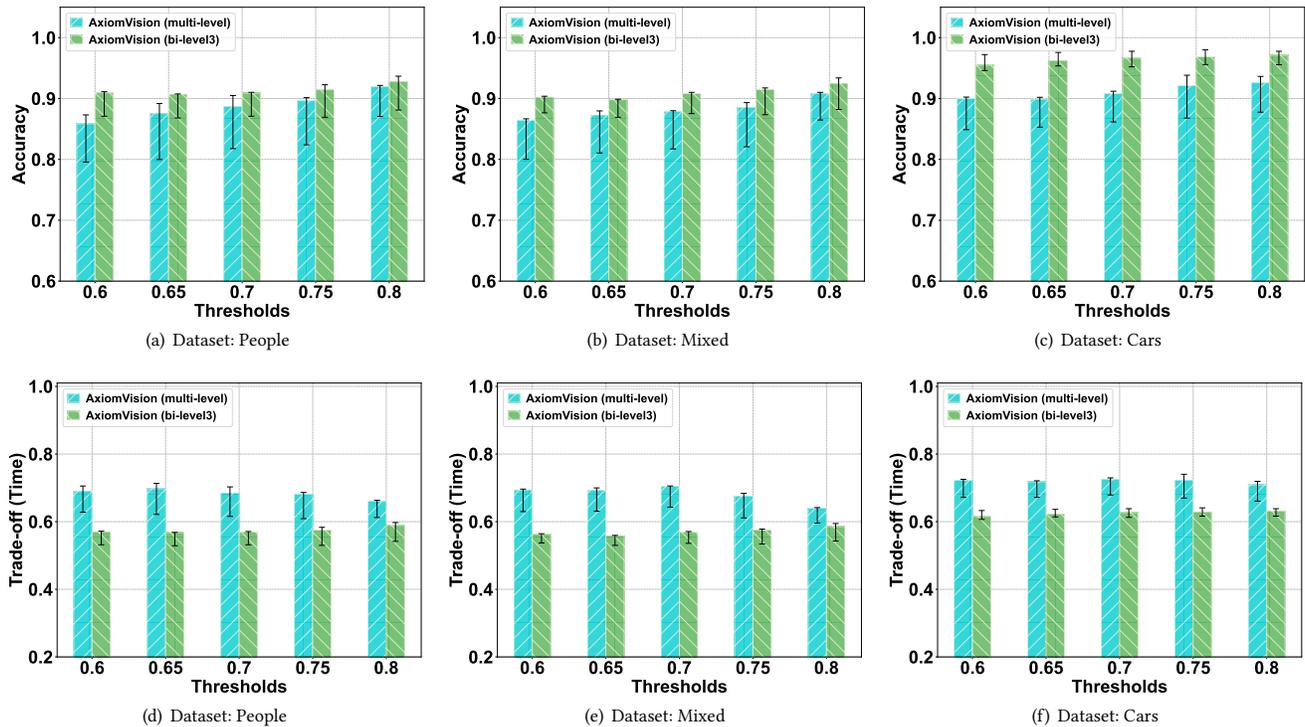


Figure 8: Comparison of *AxiomVision* with multi deployment levels; *AxiomVision* (bi-level) with the second and third levels.

tasks requiring high accuracy results in lower overall accuracy, making the overall trade-off less favorable.

Secondly, we explored another dual-level visual model deployment scheme, namely deploying only simple and complex models, designated as *AxiomVision* (bi-level2). Fig. 7(a)-(c), compared to *AxiomVision* (multi-level), *AxiomVision* (bi-level2), due to covering only simple and complex models, may more frequently invoke complex visual models, thus exhibiting higher accuracy under certain conditions. However, as shown in Fig. 7(d)-(f), *AxiomVision* (bi-level2) leads to higher resource consumption due to the omission of medium-complexity models. In contrast, *AxiomVision* (multi-level) can adopt medium-complexity models for tasks of medium complexity, ensuring ideal accuracy while efficiently saving bandwidth resources. Therefore, when evaluating the trade-off between accuracy and bandwidth, the multi-level visual model deployment strategy demonstrates significant advantages.

Finally, we examined the deployment scheme of the dual-level visual model, termed *AxiomVision* (bi-level3), which incorporates medium-complexity and complex models. As illustrated in Fig. 8(a)-(c), unlike the *AxiomVision* (multi-level) approach, *AxiomVision* (bi-level3) eschews the use of simple models and this modification enhances accuracy. However, it is worth noting that *AxiomVision* (multi-level) also achieved the threshold targets. However, when we take into account a trade-off between accuracy and execution time, denoted as $a - \gamma e$, where a and e represent normalized accuracy and time, respectively, and $\gamma = 0.2$ is a weighted unit conversion parameter. As shown in Fig. 8(d)-(f), when evaluating the trade-off

between accuracy and execution time, the multi-level visual model deployment strategy demonstrates significant advantages.

In summary, *AxiomVision* (bi-level) cannot simultaneously ensure and balance the three metrics of accuracy, bandwidth, and execution time. In contrast, *AxiomVision* (multi-level), when properly structured, can achieve better results in all three metrics. Through the discussions above, we provide deeper insights into the effective deployment of visual models, and demonstrate that the multi-level deployment strategy has outstanding adaptability to visual tasks of varying complexities—capable of maintaining high accuracy, low latency and resource consumption, presenting an optimal solution.