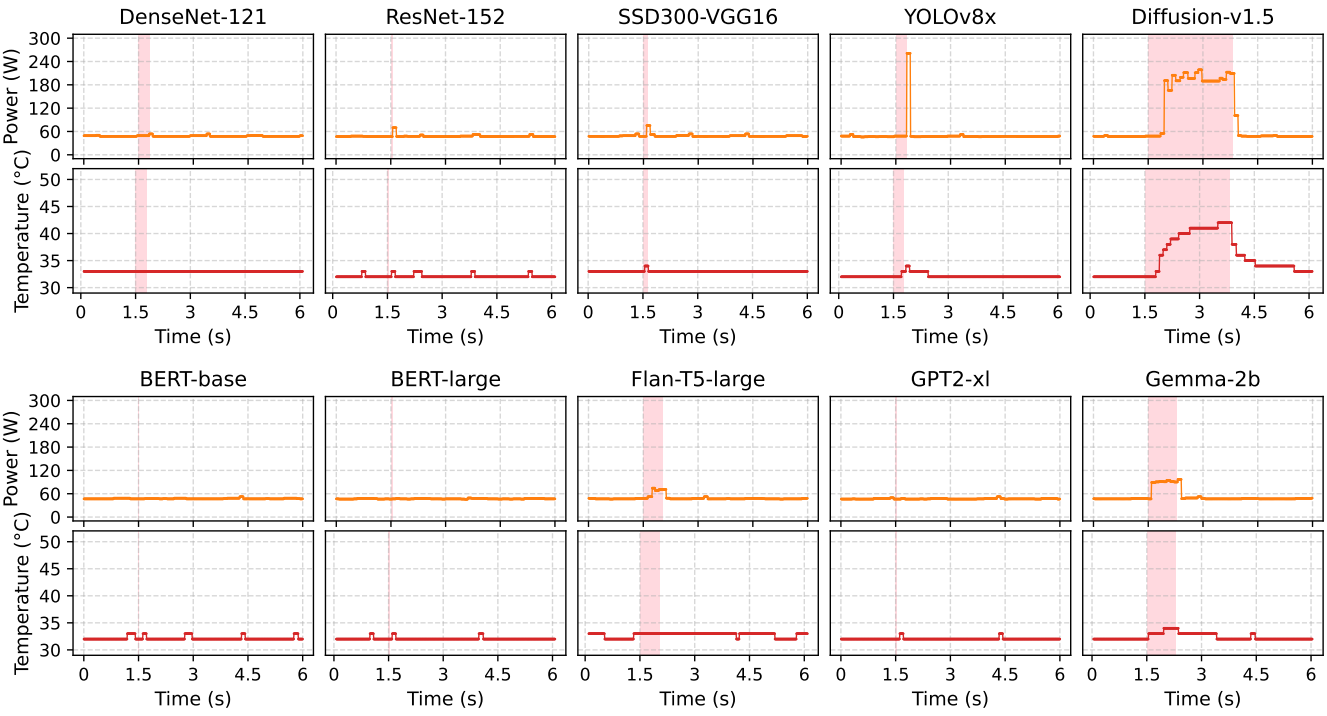
InferCool: Enhancing AI Inference Cooling through Transparent, Non-Intrusive Task Reassignment

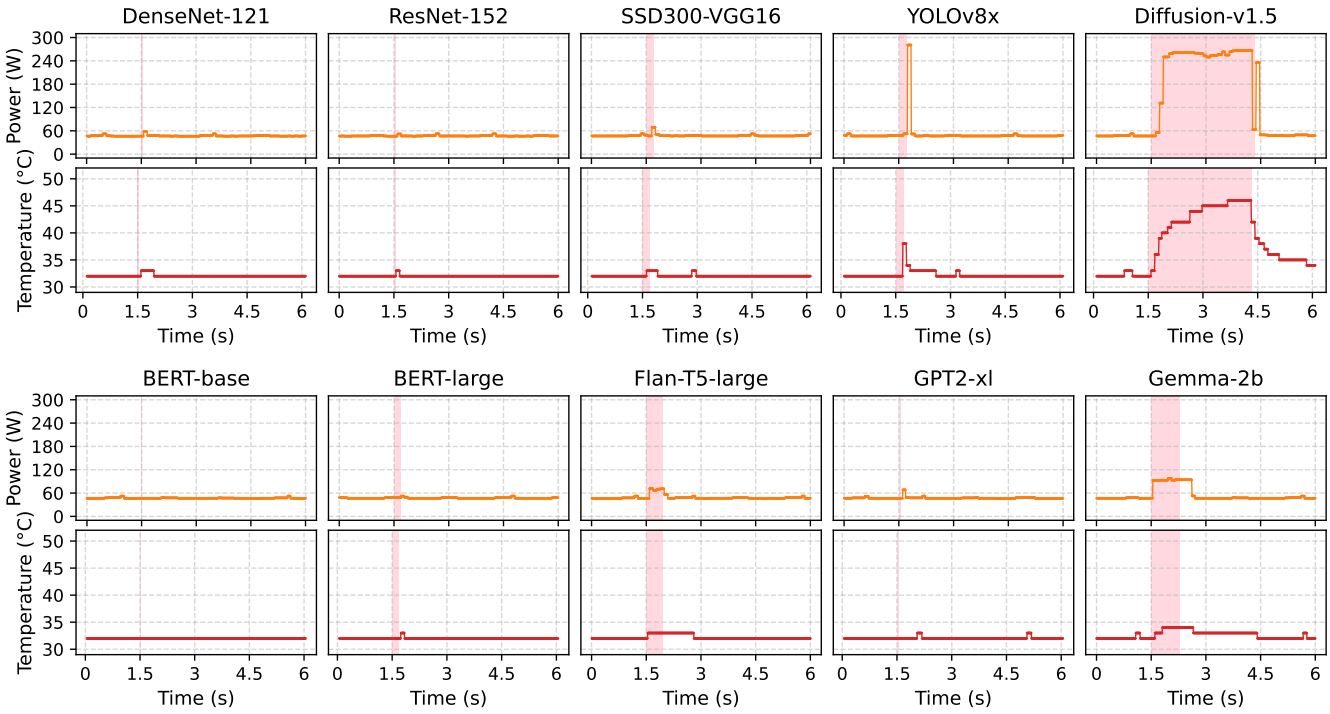
Supplementary File

Qiangyu Pei
Huazhong University of Science and Technology
Wuhan, China
peiqiangyu@hust.edu.cn

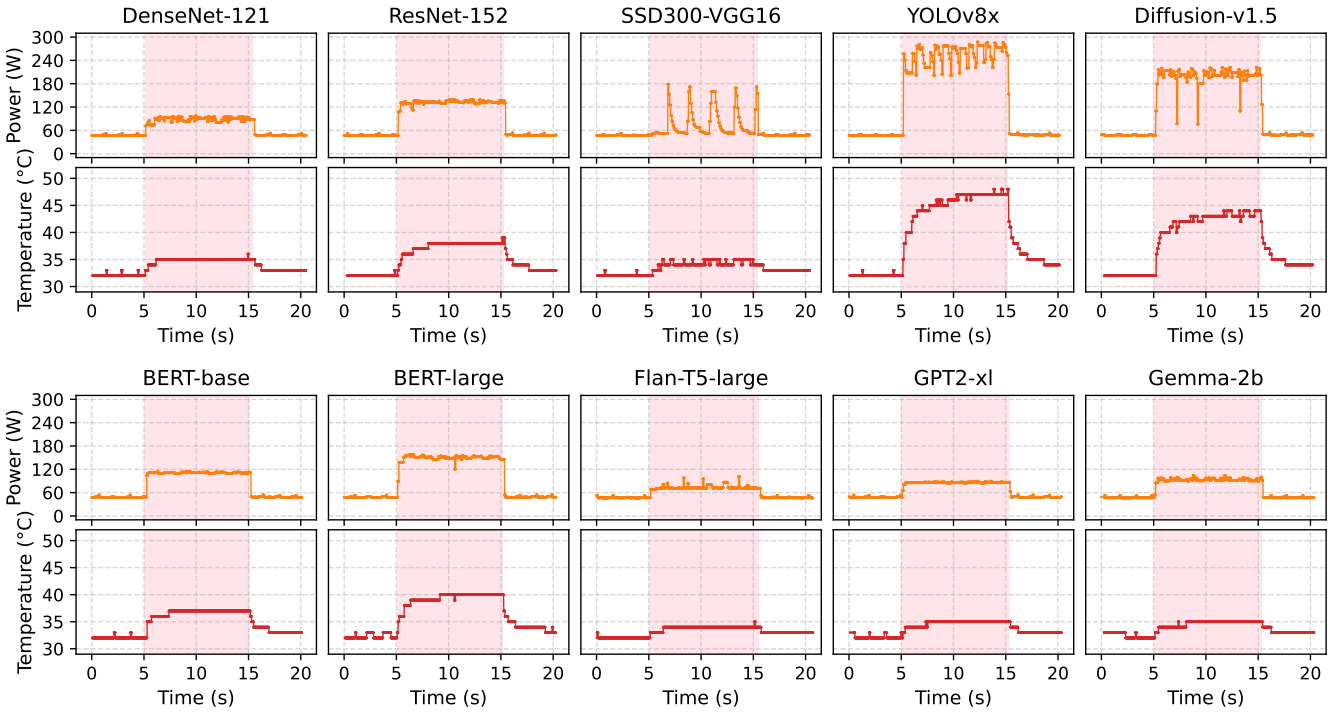


(a) Running inference *only once (bs = 8)*

Figure 3: The GPU power and temperature variations under different individual and cumulative intensities.

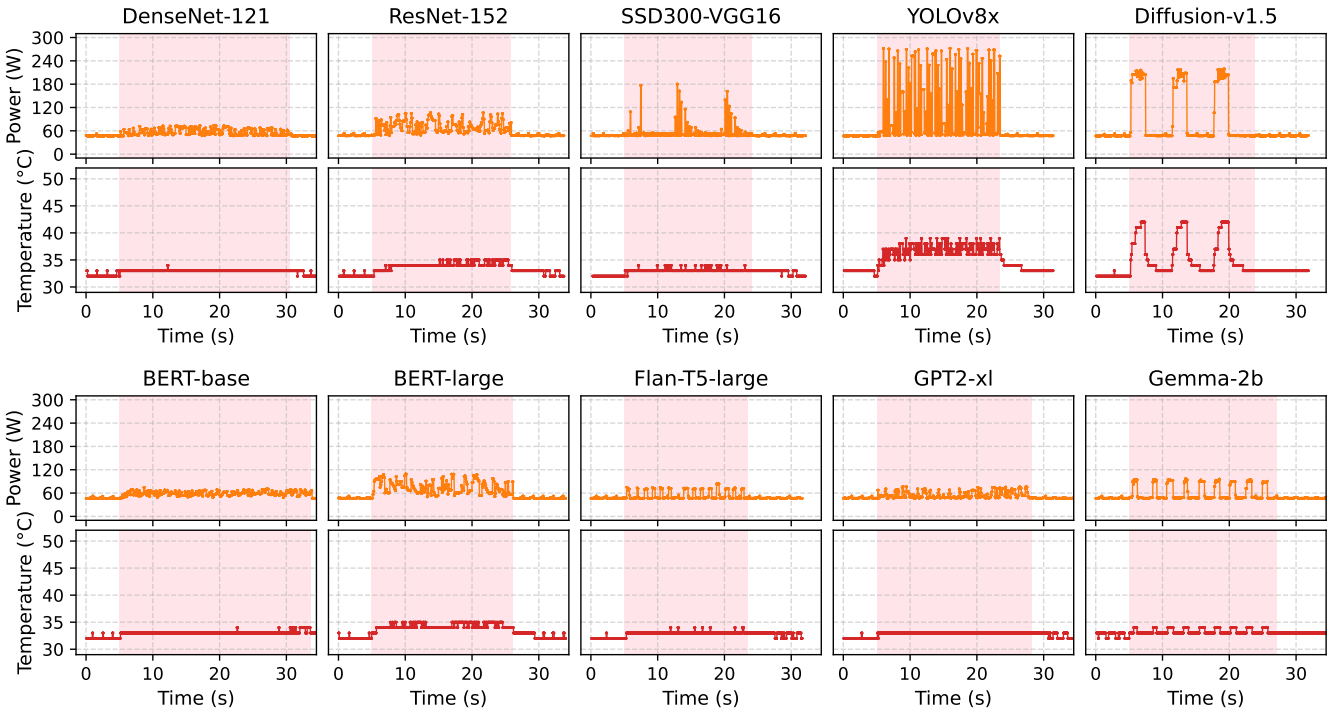


(b) Running inference *only once (bs = 16)*

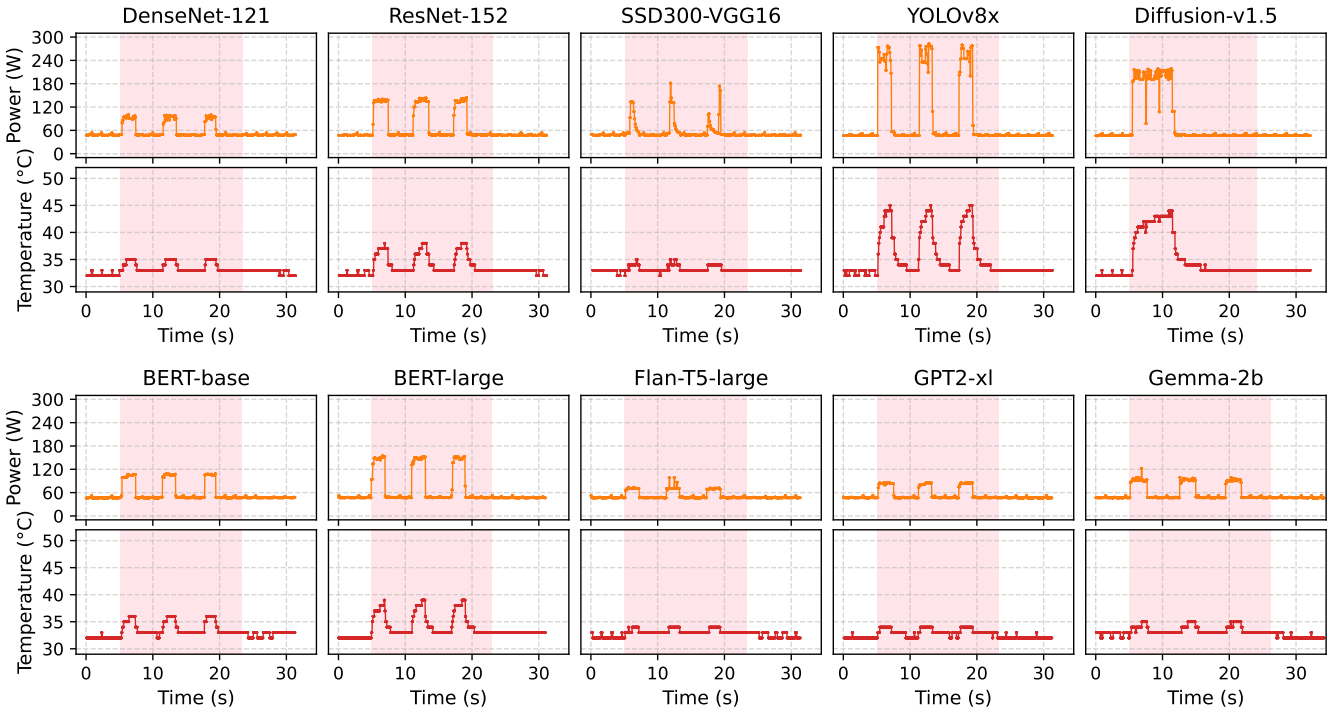


(c) Running inference *continuously for 10 seconds (bs = 8)*

Figure 3: The GPU power and temperature variations under different individual and cumulative intensities.

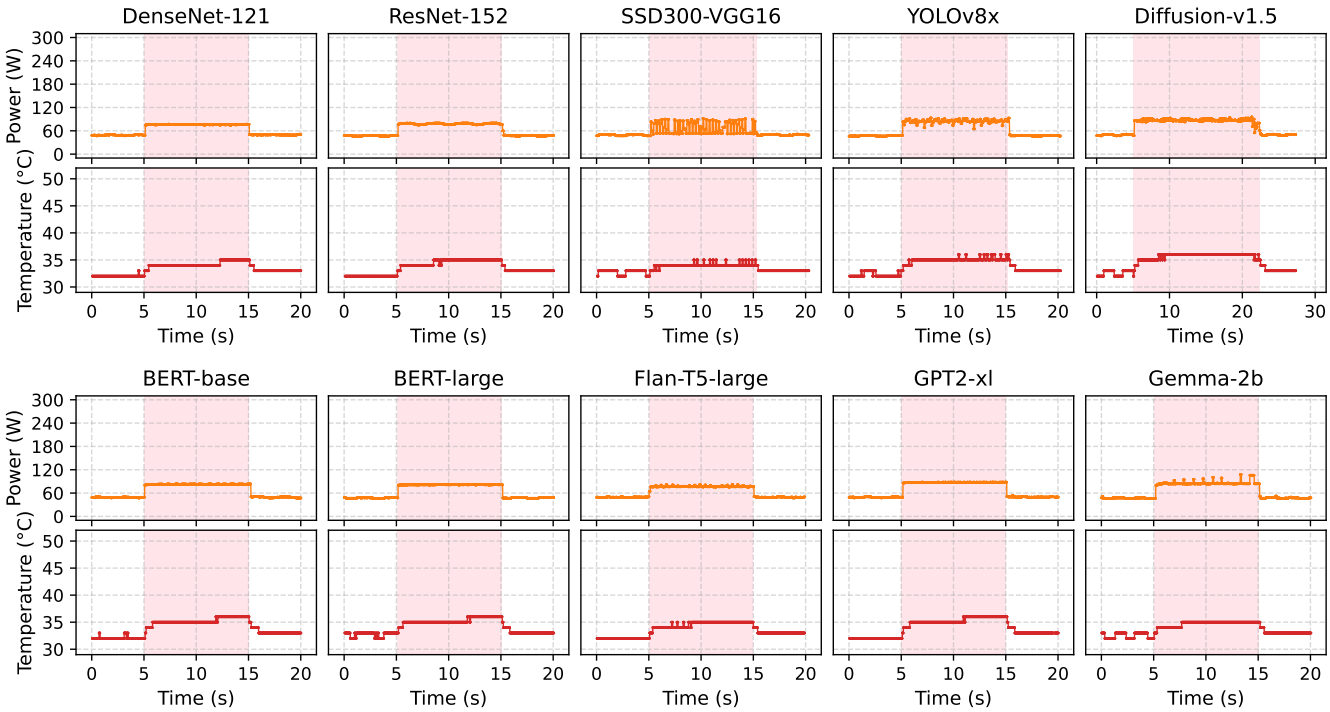


(a) ***Uniform request arrival:*** **after each inference execution, stopping for twice the inference time (bs = 8)**

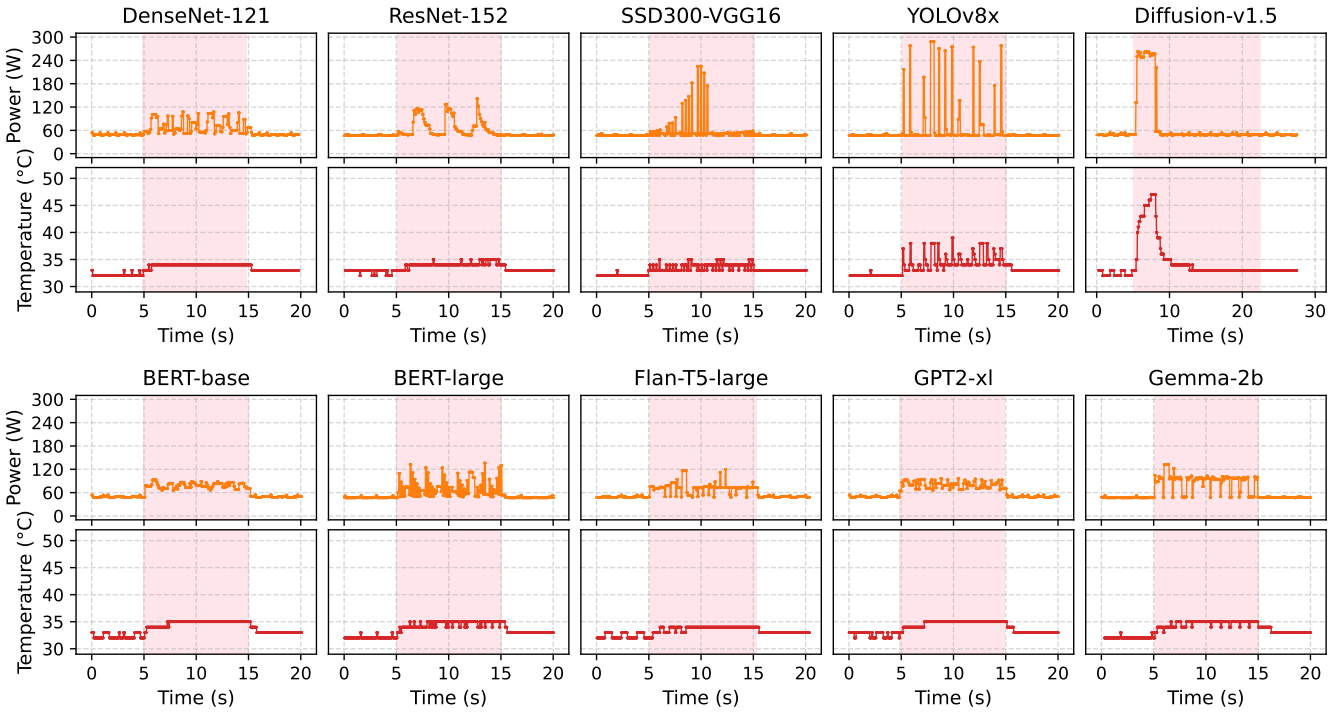


(b) ***Bursty request arrival:*** **for every 2 seconds of inference executions, stopping for 4 seconds (bs = 8)**

Figure 4: The GPU power and temperature variations under different intensity distributions.

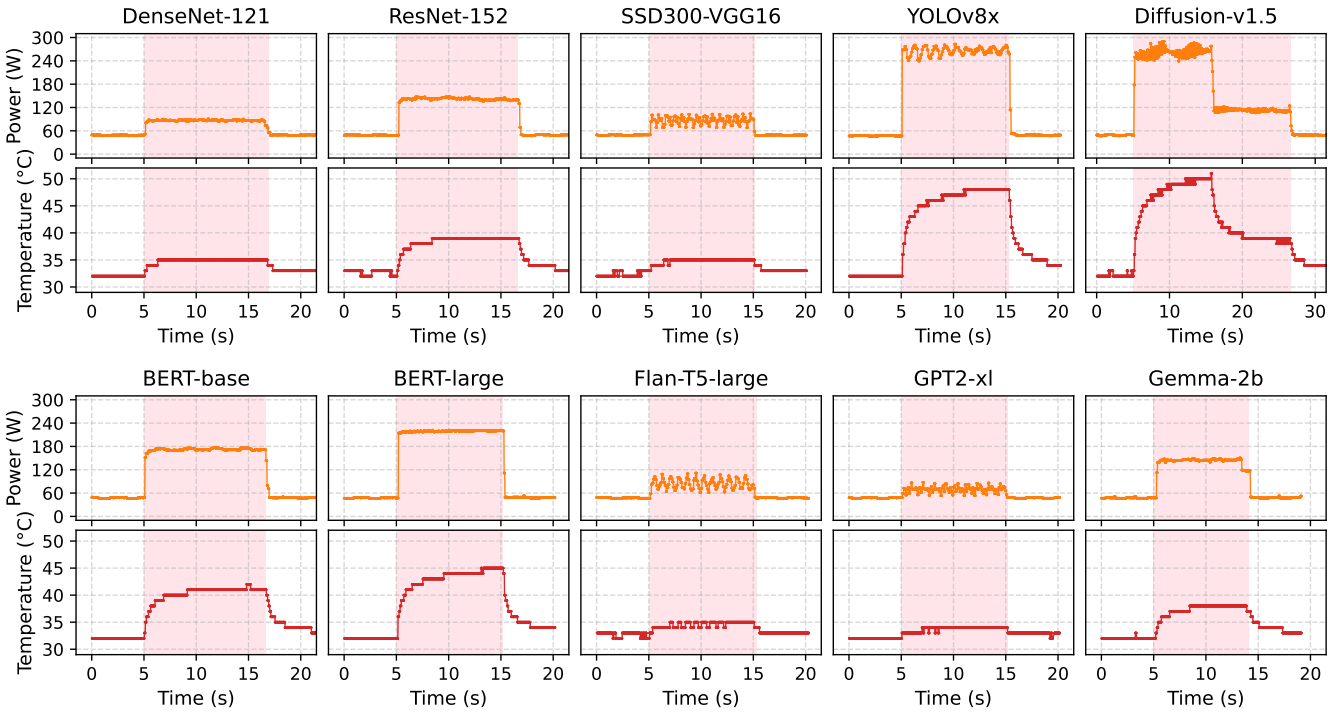


(a) *Smaller partition:* one `1g.5gb` partition (bs = 16)

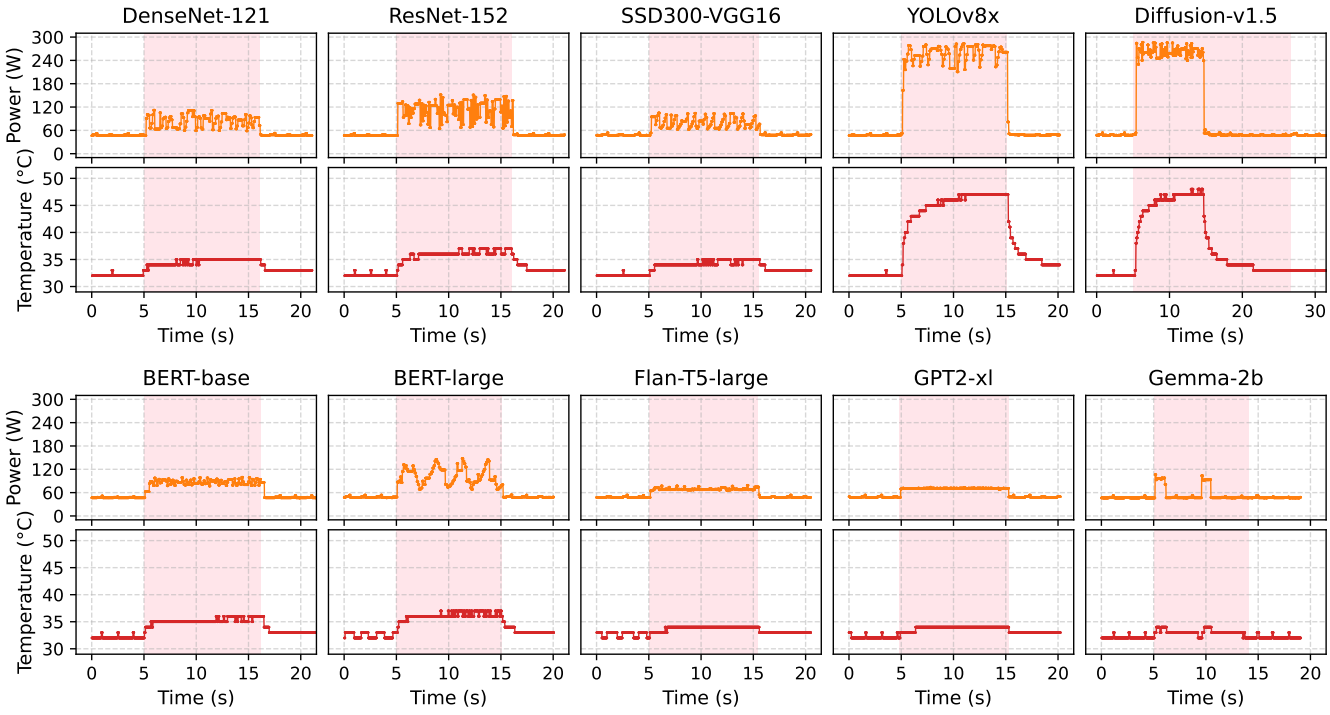


(b) *Larger partition:* one `7g.40gb` partition (bs = 16)

Figure 5: The GPU power and temperature variations under different cumulative intensities influenced by GPU partitioning plans.

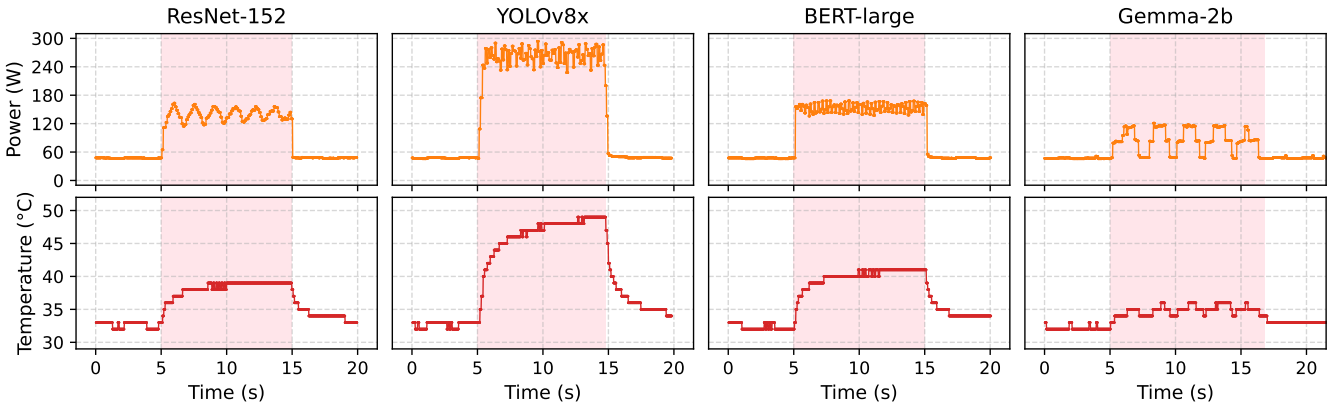


(a) *Smaller partition:* bs = 1 on seven `1g.5gb` partitions

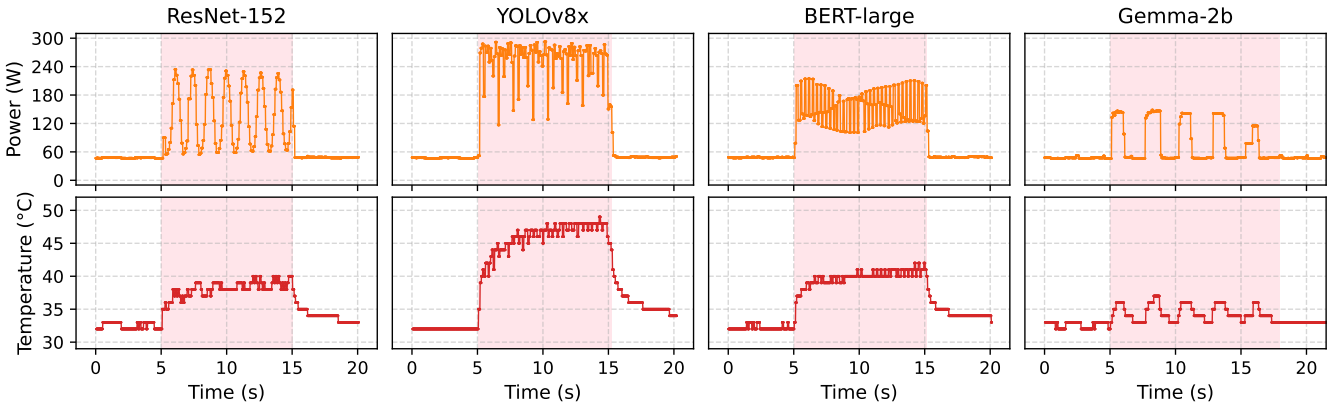


(b) *Larger partition: Adaptive bs* on one `7g.40gb` partition

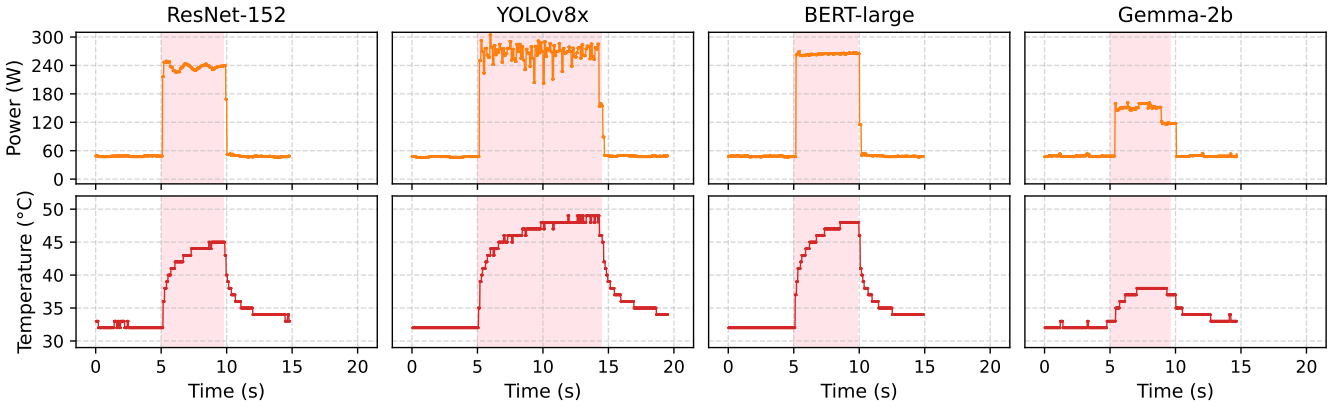
Figure 6: The GPU power and temperature variations under different cumulative intensities influenced by GPU partitioning plans and batch sizes.



(a) *Uniform request arrival:* requests arrive uniformly over the time period

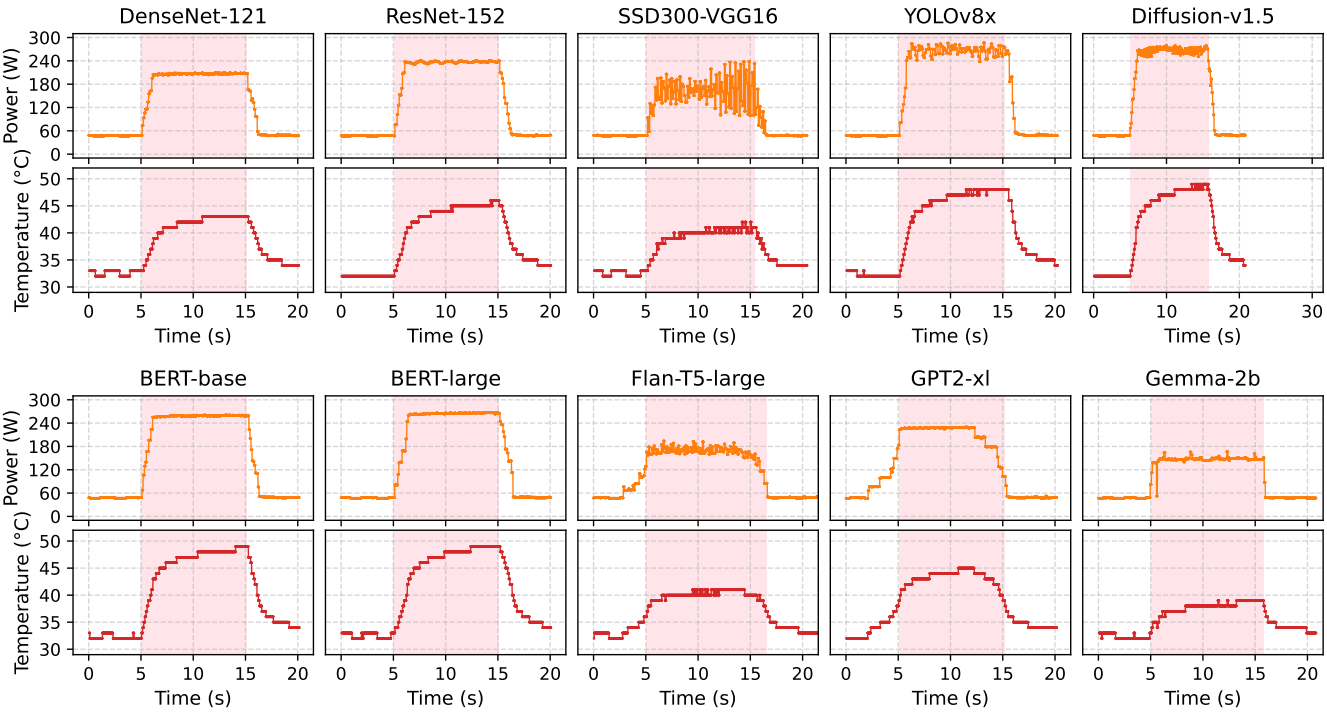


(b) *Small bursty request arrival:* requests cluster at the beginning of each time slice



(c) *Large bursty request arrival:* all requests arrive at the beginning of the time period

Figure 16: The GPU power and temperature variations when running inference on seven `1g.5gb` partitions simultaneously under different arrival patterns and under full load (bs = 8). In Figures (a), (b), and (c), the amount of requests is the same as that in Figure 3c.



(d) *Full load:* Requests arrival continuously over the time period

Figure 16: The GPU power and temperature variations when running inference on seven `1g.5gb` partitions simultaneously under different arrival patterns and under full load (bs = 8). In Figures (a), (b), and (c), the amount of requests is the same as that in Figure 3c.

A PROOF FOR THE OBSERVATION

Here, we provide a proof for the observation stated in Section 5.3: the sum of the temperature increases across all partitions is approximately equal to the overall temperature increase of the GPU component.

Proof. Assume there are m models co-located on the GPU, with their inference powers denoted as $P_j, j \in \{1, 2, \ldots, m\}$. The steady-state GPU temperatures when executing each model individually are $T_j, j \in \{1, 2, \ldots, m\}$. Let P_{total} and T_{total} represent the total GPU power and steady-state GPU temperature when executing all models concurrently. The idle-state GPU power and temperature are denoted as P^S and T^S, respectively. Based on the Equations 1, 3, and 8 from the paper:

$$P^{G_i} = \sum_{j,\, M_j \in G_i} P_{j,k,b} - (|G_i| - 1) \cdot P^S, \tag{1}$$

$$P^{G_i}_{\text{steady}} = hA(T^{G_i}_{\text{steady}} - T_w), \tag{2}$$

$$T(t) = (T_0 - T^{G_i}_{\text{steady}'})e^{-\alpha t} + T^{G_i}_{\text{steady}'}, \tag{3}$$

we can get:

$$T^S = T_w + \frac{P^S}{hA}, \tag{4}$$

$$T_j = T_w + \frac{P_j}{hA}, \quad j \in \{1, 2, \ldots, m\}, \tag{5}$$

$$T_{\text{total}} = T_w + \frac{P_{\text{total}}}{hA}, \tag{6}$$

$$P_{\text{total}} = \sum_{j=1}^{m} P_j - (m-1) \cdot P^S. \tag{7}$$

Thus, we can derive the relationship between the temperature increase of the GPU component and the sum of temperature increases of all partitions at the steady state:

$$\begin{aligned}(T_{\text{total}} - T^S) &= \frac{P_{\text{total}}}{hA} - \frac{P^S}{hA} \\ &= \sum_{j=1}^{m} \frac{P_j}{hA} - (m-1) \cdot \frac{P^S}{hA} - \frac{P^S}{hA} \\ &= \sum_{j=1}^{m} \left(\frac{P_j}{hA} - \frac{P^S}{hA} \right) \\ &= \sum_{j=1}^{m} \left(T_j - T^S \right). \end{aligned} \tag{8}$$

Equation 8 shows that at the steady state, the overall temperature increase of the GPU component is approximately equal to the sum of temperature increases of all partitions.

At the non-steady state, let T_0 represent the current GPU temperature. When $T_0 = T^S$, based on Equations 3 and 8, we have:

$$\begin{aligned}T_{\text{total}}(t) - T^S &= T_{\text{total}} + (T^S - T_{\text{total}})e^{-\alpha t} - T^S \\ &= (T_{\text{total}} - T^S)(1 - e^{-\alpha t}) \\ &= \sum_{j=1}^{m} (T_j - T^S)(1 - e^{-\alpha t}) \\ &= \sum_{j=1}^{m} \left(T_j(t) - T^S \right). \end{aligned} \tag{9}$$

When T_0 is above T^S, i.e., from a non-steady state, we have:

$$\begin{aligned}T_{\text{total}}(t) - T^S &= T_{\text{total}} + (T_0 - T_{\text{total}})e^{-\alpha t} - T^S \\ &= (T_{\text{total}} - T^S)(1 - e^{-\alpha t}) + (T_0 - T^S)e^{-\alpha t} \\ &= \sum_{j=1}^{m} (T_j - T^S)(1 - e^{-\alpha t}) + (T_0 - T^S)e^{-\alpha t} \\ &= \sum_{j=1}^{m} (T_j(t) - T^S) + \left((T_0 - T^S) - \left(\sum_{j=1}^{m} (T_{j,0} - T^S) \right) \right) e^{-\alpha t}. \end{aligned} \tag{10}$$

From Equation 9, we can know $(T_0 - T^S) = (\sum_{j=1}^{m} (T_{j,0} - T^S))$. Therefore, Equation 10 can be further written as:

$$T_{\text{total}}(t) - T^S = \sum_{j=1}^{m} (T_j(t) - T^S) \tag{11}$$

Equation 11 shows that, in a non-steady state, the overall temperature increase of the GPU component is still approximately equal to the sum of temperature increases of all partitions. □