Attention-Driven Gaussian Modeling for Total Duration of Heterogeneous Operations

J. Edgar Hernandez

Motivation & Objective

Motivation: Many industrial tasks can be viewed as composite operations made up of elementary steps. Having a reliable estimator for the "standard" duration of such operations enables objective, data-driven assessment of performance—quantifying efficiency gains or losses whenever operational policies or staffing changes.

Objective: Build a predictive baseline which, given descriptive features of an operation and its sub-operations (e.g. dimensions, weight, count, crew size), can:

1. **Estimate the typical duration** $\hat{T}$ of any composite operation, and associated uncertainty $\sigma_{\hat{T}}$.
2. **Act as a performance benchmark** by comparing observed times T_{obs} to $\hat{T}$, flagging deviations and process bottlenecks.

Case study: trailer loading in a distribution plant, where each bulk operation moves a set of distinct products into a trailer by hand.

Operation Structure

An *operation* is any task decomposable into N *sub-operations*. We represent:

$$\text{Operation} = \Big\{ \underbrace{\mathbf{x}_{\text{op}}}_{\substack{\text{global}\\\text{features}}},\ \{ \underbrace{\mathbf{x}_{\text{sub},i}}_{\substack{\text{item-level}\\\text{features}}} \}_{i=1}^{N} \Big\}$$

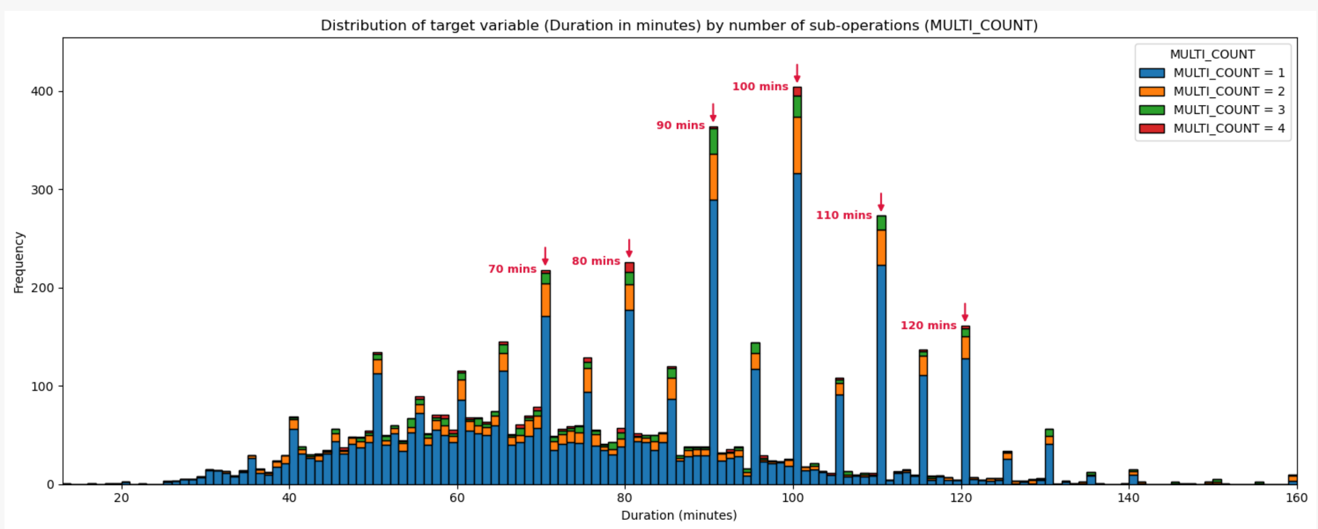
Operation-Level Features $\mathbf{x}_{\text{op}}$: crew size, total load volume, etc. **Sub-Operation Features $\mathbf{x}_{\text{sub},i}$:** for each product—dimensions, weight, quantity, handling category.

In our trailer-loading case, each sub-operation is one product in the shipment.

Biased Duration Distribution

The operation durations in our dataset were recorded manually, introducing systematic bias. Instead of a smooth bell curve, the empirical distribution shows:

- **Rounding artifacts**—operators tend to report "nice" round times (e.g. large spikes at 90 or 100 min) rather than the exact duration.
- **Extreme variability**—occasional mis-entries and outliers produce heavy tails and long right-hand skew.



The heavy-tailed distribution with pronounced rounding peaks requires a modeling approach that's robust to noise and outliers, but that is equally representative of normal operations, and short or long under-represented durations.

Furthermore, the dataset is highly imbalanced in terms of sub-operations: single sub-operation cases dominate, while those with 2, 3, or 4 sub-operations are much rarer:

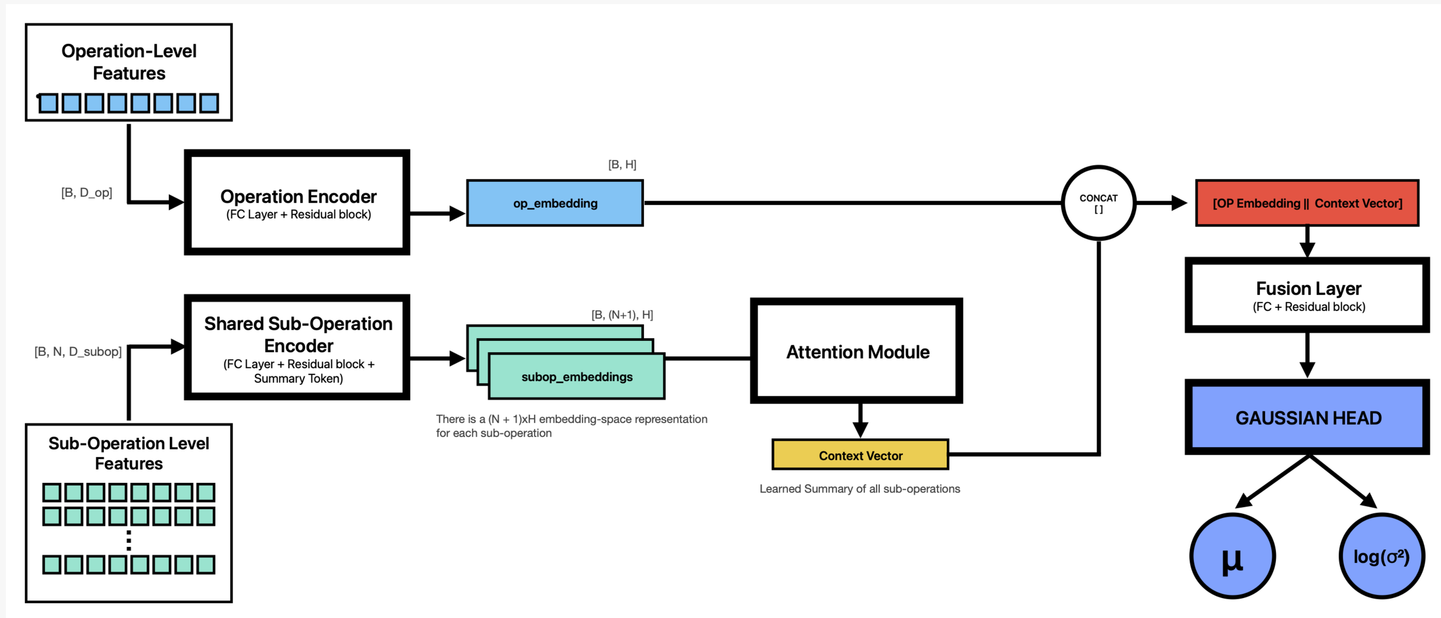
Num. of Sub-Operations	1	2	3	4
Percentage	78.0%	14.1%	6.0%	2.0%

Moreover, the scarcity of operations with higher sub-operation counts calls for upsampling and data-augmentation strategies to ensure effective training.

Architecture

Our `DurationPredictor` follows a two-stream encode–attend–fuse pipeline:

- **Operation Encoder:** A linear projection of the global operation features into a D-dim latent space, followed by a ResidualBlock to learn small refinements on top of the identity path.
- **Sub-Operation Encoder:** Each sub-operation vector (plus a learned "summary" token) is projected and refined via the same ResidualBlock, yielding per-token embeddings of shape $(B, N{+}1, D)$.
- **Attention Module:** An MLP scores each sub-op embedding, applies softmax to get α_i, and computes a weighted sum $\sum_i \alpha_i h_i$ as the context vector.
- **Fusion Block:** We concatenate the operation embedding and the context, project back to D dimensions, and apply another ResidualBlock to merge global and local signals.
- **Heteroskedastic Gaussian Head:** A final MLP predicts both mean $\mu(x)$ and scale $\sigma(x) > 0$, trained with a loss robust to outliers that adapts to input-dependent uncertainty.



Model flow and architecture

The loss function

We optimize a composite loss combining a heteroskedastic Gaussian negative-log-likelihood, a median-quantile bias term, and a slope penalty:

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{NLL}}}_{\substack{\text{heteroskedastic}\\\text{Gaussian NLL}}} + \beta\ \underbrace{\mathcal{L}_{\text{quantile}}^{\tau=0.5}}_{\text{median bias}} + \gamma\ \underbrace{\mathcal{L}_{\text{slope}}}_{\substack{\text{slope}\\\text{penalty}}}$$

1. Heteroskedastic Gaussian NLL:

$$\mathcal{L}_{\text{NLL}} = \frac{1}{B}\sum_{i=1}^{B}\Big[\frac{(y_i-\mu(x_i))^2}{2\,\sigma(x_i)^2} + \frac{1}{2}\log\big(2\pi\,\sigma(x_i)^2\big)\Big],$$

where the network predicts both $\mu(x)$ and $\sigma(x) > 0$ for each example.

2. Median-quantile (pinball) loss:

$$\mathcal{L}_{\text{quantile}}^{\tau} = \frac{1}{B}\sum_{i=1}^{B}\rho_\tau\big(y_i - q_\tau(x_i)\big), \qquad \rho_\tau(u) = \begin{cases}\tau\,u, & u \ge 0,\\ (\tau-1)\,u, & u < 0,\end{cases}$$

with $\tau = 0.5$ to penalize median bias.

3. Slope penalty: Compute weighted residuals $r_i = \hat{y}_i - y_i$ (downweighting top 10% outliers) and centered targets $y_i' = y_i - \bar{y}$. Then

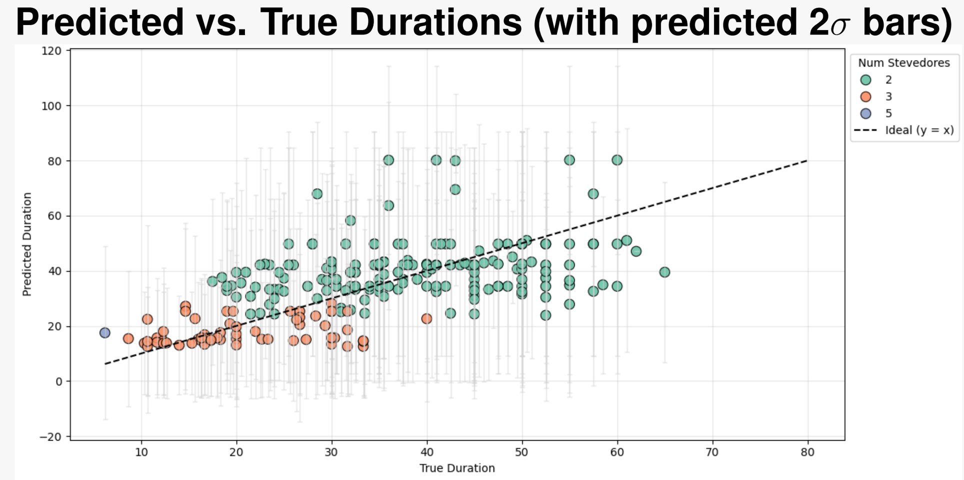
$$\text{slope} = \frac{\sum_i w_i\, r_i\, y_i'}{\sum_i w_i\,(y_i')^2}, \qquad \mathcal{L}_{\text{slope}} = (\text{slope})^2.$$

This discourages any linear trend between residuals and true values.

Prediction Results on Test Data

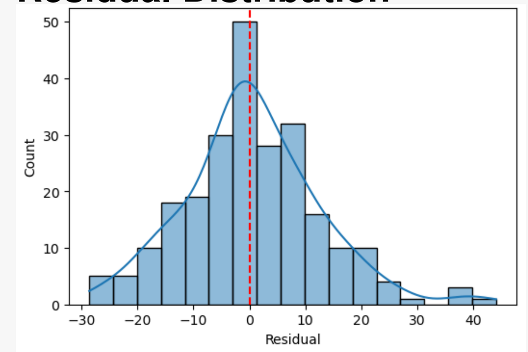
Below we compare model forecasts against actual load durations and summarize key performance metrics.

Predicted vs. True Durations (with predicted 2σ bars)



Scatter of predicted $\hat{T}$ against observed T_{obs}.

Residual Distribution



Histogram of residuals $\hat{T} - T_{\text{obs}}$, showing most errors near zero and a few outliers.

Summary Metrics:

90% CI coverage	Median bias	Mean residual (bias)	MAE	MAPE
88.4%	-0.32	0.51	9.05	29.34%

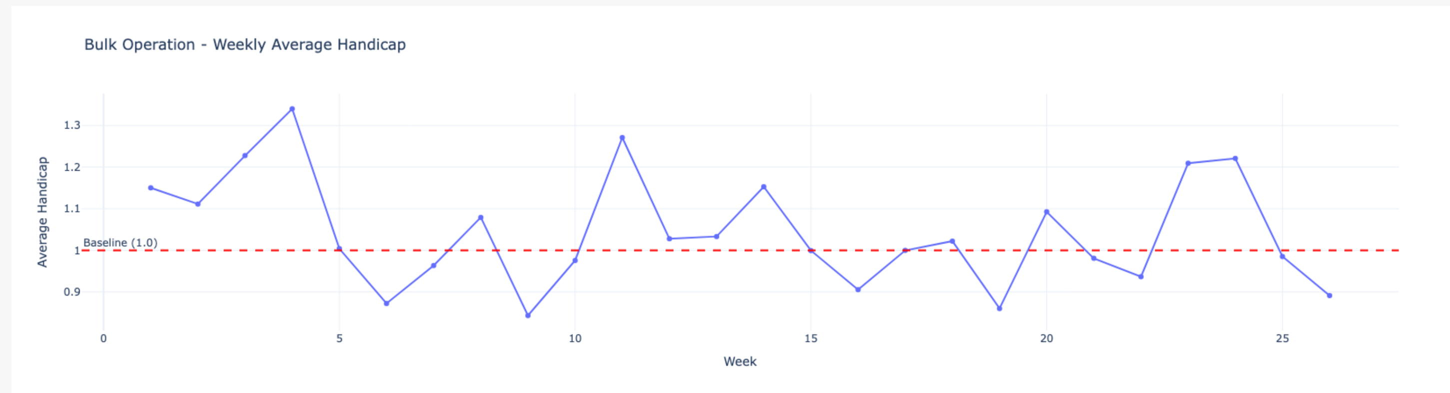
Performance Estimation

We quantify operational performance via the *Handicap* metric, defined for each operation i as

$$\text{Handicap}_i = \frac{T_i^{\text{capped}}}{\hat{T}_i} \quad \text{where} \quad T_i^{\text{capped}} = \min\big(T_i,\ \hat{T}_i + 2\,\sigma_i\big).$$

Here $\hat{T}_i$ is the model's baseline prediction and σ_i its estimated uncertainty; capping at $\hat{T}_i + 2\sigma_i$ limits outlier influence.

The model was trained on historical 2024-25 data and fine-tuned for 2025 loading dynamics. The goal was to center Handicap distribution around 1.0, ensuring the baseline accurately reflects 2025's "usual" performance.



Aggregate 2025 results: Mean Handicap $= \mathbf{1.05}$; Median Handicap $= \mathbf{1.01}$.

Discussion

- **Benchmarking:** Our estimator with handicap centered around 1.01 serves as a reliable baseline for trailer-loading performance; operations deviating above or below flag inefficiencies or best practices.
- **Robustness:** The heteroskedastic Gaussian NLL with bias/slope penalties effectively copes with manual-entry spikes and imbalanced multi-SKU loads.
- **Future work:**
  - Collect more real-world examples of multi-SKU loads to better represent higher-order sub-operations; if data remains sparse, generative models could be used to synthesize realistic samples in under-represented regions.
  - Adapt the architecture to model the *order* of sub-operations and capture any sequential dependencies in loading steps.
  - Incorporate additional contextual features (shift timing, equipment availability, other conditions) so the estimator can adjust for different operational variability.

edgarcancinoe@outlook.com