DaVoS: An Instantaneous Safety Model for Autonomous Vehicles

Yuankai He, Yuxin Wang, Boyang Tian, Weisong Shi
University of Delaware, Newark, USA
{willhe, yuxw, tby, weisong}@udel.edu

***Abstract*—Reliable, interpretable measurement of autonomous vehicle (AV) safety remains a key obstacle to widespread deployment. In practice, industry-facing summaries emphasize exposure-normalized outcomes (e.g., crashes per mile), while research metrics such as time-to-collision (TTC) provide scenario-level kinematic diagnostics; both are valuable but do not yield a single physics-grounded scalar that exposes the coupling between predicted conflicts and the distance traveled during nonzero sense–decide–act latency. This paper proposes DaVoS, a physics-based, platform-agnostic model for evaluation: an instantaneous formulation that summarizes safety confidence under the current plan and predictions using five measurable quantities—distance to first predicted conflict along the ego planned trajectory, current speed, longitudinal acceleration/deceleration, end-to-end sense–decide–act overhead, and a variability descriptor that selects conservative derating. The DaVoS model allows a vehicle to calculate a safety-confidence score at any instantaneous time from these metrics. DaVoS compares conservative available distance to conservative committed distance over the overhead window and maps the resulting slack to a bounded score. The metric is intended for comparison on matched scenario sets under a disclosed conflict predicate, horizon, conservative policy, and score calibration. We illustrate the score on two highway scenarios (car-following and an adjacent-lane cut-in) and provide measurement guidance for reproducible reporting.**

I. Introduction

Measuring autonomous vehicle (AV) safety remains a central obstacle to widespread deployment. Exposure-normalized outcome statistics (e.g., crashes or interventions per X miles) are intuitive but dominated by rare events [1]. Scenario-level proxies such as time-to-collision (TTC) and time headway are diagnostically useful [2], [3], but they often omit end-to-end latency. These gaps leave developers without a single physics-grounded *system-level* scalar that is sensitive to both physical margins and timing overhead.

This paper introduces DaVoS, an instantaneous, platform-agnostic safety-confidence model. DaVoS computes an instantaneous score from five measurable quantities: distance to first predicted conflict along the planned ego trajectory, current speed, longitudinal acceleration/deceleration, end-to-end sense–decide–act overhead, and a variability descriptor that selects conservative derating. By comparing conservative distance-to-conflict with conservative distance committed during the overhead window, score changes localize low confidence to prediction/planning, motion policy, timing overhead, or conservatism.

DaVoS is not an absolute safety certificate; it is a compact, physically interpretable *comparative* reference when the scenario set and conservative policy are fixed and disclosed. The model makes explicit the coupling between plan-conditioned predicted conflicts and the distance traveled during nonzero end-to-end latency.

We define the model and its required inputs, provide a calibrated score mapping, and illustrate interpretation on representative highway scenarios. The remainder of the paper motivates the need for such a reference (Section II), defines the DaVoS model (Section III), and demonstrates interpretation on illustrative scenarios.

II. Background and Motivation

Deployment of autonomous vehicles requires sustained public and regulatory trust, yet safety is still summarized through heterogeneous artifacts that are difficult to compare. Existing signals span exposure-normalized outcomes, component-level learning scores, and assurance evidence, but they do not provide a compact physical quantity that characterizes near-term safety confidence under a system's current plan and predictions [4].

In current practice, organizations assemble layer-specific measurements into a safety case. Outcome statistics are indispensable but dominated by rare events, requiring substantial exposure [1]. Disengagement and intervention signals depend on intervention policy and reporting conventions [5], [6], and formats vary by jurisdiction [7]. Component-level perception and prediction metrics quantify model fidelity but do not compose into system-level physical margins under nonzero end-to-end delay.

Planning- and driving-quality proxies such as TTC, time headway, and jerk provide useful diagnostics [2], [3], but are scenario dependent and often rely on implicit reaction assumptions. Assurance frameworks emphasize structured evidence across hardware, software, and learned components [8]–[10].

Taken together, these approaches leave a key coupling implicit: near-term safety confidence depends on what the system predicts will occur along its planned motion and on how long it takes the system to enact a materially different decision while the vehicle continues to move. Many physics-flavored proxies are computed from instantaneous geometry or under an implicit assumption of immediate reaction; in practice, the question is whether the system has sufficient conservative headroom given nonzero delay.

This motivates a compact physics-grounded model—expressed in terms of distance, speed, acceleration, and end-

to-end delay—that can be computed consistently across stacks and incorporates measured variability through a disclosed conservative policy. DaVoS fills this need by reducing the system to five measurable quantities and mapping them to a monotone instantaneous score. By construction, DaVoS makes the prediction–latency coupling explicit in a single physically interpretable scalar.

A. Related Work

DaVoS is related to TTC/headway and safe-distance models that incorporate reaction delay and braking limits [2], [3], [11]. It is also related to formal safety-envelope approaches such as RSS and its treatment of variability [12], [13], and to safety standards that emphasize structured evidence [8]–[10]. In contrast to these artifacts, DaVoS is defined as a plan-conditioned, latency-explicit instantaneous score and is intended for consistent comparison under a disclosed metric specification.

III. The DaVoS Model

DaVoS is an instantaneous, platform-agnostic model that maps five measurable quantities to a scalar safety-confidence score. The model is physics-based in that it compares (i) an available distance-to-conflict along the planned ego motion with (ii) a committed distance traveled during measured end-to-end latency, computed from speed and acceleration. At time t_0, DaVoS compares the conservative distance remaining until the earliest predicted conflict along the current planned ego motion with the conservative distance the ego is expected to travel before a new decision can take effect. The five quantities are: distance-to-first-predicted-conflict along the planned ego trajectory ($D^\star$), current ego acceleration/deceleration (a), measured variability (V), end-to-end sense–decide–act overhead (o), and current ego speed (S).

A. What gap the model fills

As motivated in Section II, existing AV evaluation signals are informative but heterogeneous and do not yield a single standardized scalar that couples plan-conditioned predicted conflicts with the distance traveled during nonzero end-to-end latency. DaVoS fills this gap by defining a physics-grounded instantaneous quantity in terms of $(D^\star, S, a, o, V)$, together with disclosed conservative transforms that map measured/computed quantities to conservative counterparts such as $D_V^{\downarrow}$ and $o_V^{\uparrow}$. The instantaneous score defined by the model can be used directly for online monitoring, plotted as a time series over a scenario, or aggregated into scenario-level summaries (Section V).

DaVoS is intended as a *comparative* metric rather than an absolute safety certificate. Comparisons across stacks or compute platforms are meaningful only when the scenario set, conflict predicate $\mathcal{C}_i$, $(T^{plan}, \Delta t)$, measurement definitions for (S, a, o), conservative transforms $\mathcal{Q}_V^{(\cdot)}$, and score calibration (σ, d_0) are matched and disclosed. If scenario-level aggregation is reported, the aggregation rule must also be disclosed.

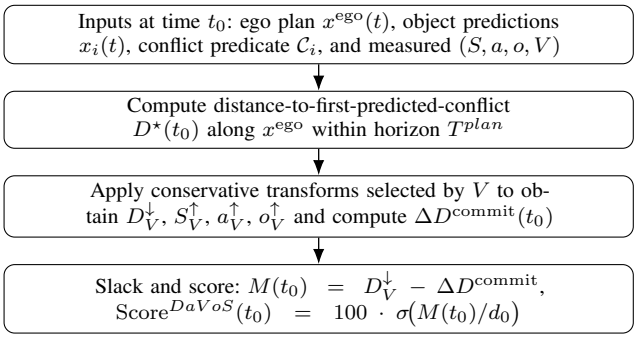

Fig. 1. DaVoS model overview. The score is determined by conservative distance-to-conflict and conservative committed distance over the end-to-end overhead window, with variability V selecting the conservative policy.

B. Notation

Table I summarizes the notation used throughout the paper. Unless otherwise stated, quantities are defined at the current time t_0 and units are SI. In particular, T^{plan} denotes the planning/prediction horizon and Δt the sampling resolution used when searching for predicted conflicts along the ego plan.

C. Scene representation and conflict predicate

At the current time t_0, let the planned ego trajectory be $x^{\mathrm{ego}}(t)$ for $t \geq t_0$, and let each object $i \in \{1, \ldots, N\}$ have a predicted trajectory $x_i(t)$ for $t \geq t_0$, defined over a finite horizon.

Define a conflict predicate for each object i:

$$\mathcal{C}_i\big(x^{\mathrm{ego}}(t), x_i(t)\big) \in \{\mathrm{true}, \mathrm{false}\}.$$

$\mathcal{C}_i$ is true when ego and object i violate the safety boundary that the model intends to protect. For example, $\mathcal{C}_i$ may indicate buffered footprint overlap, violation of a minimum separation rule, or contact with forbidden space (lane boundary, curb, non-drivable region) treated as a special "object." The exact definition is an input to DaVoS; the model requires only that the predicate can be evaluated from the plan and predictions.

D. Distance to first predicted conflict along the planned ego motion ($D^\star$)

All DaVoS quantities below are defined at the current time t_0.

For each object i, define the earliest predicted conflict time

$$t_i^\star = \inf\left\{t \geq t_0 : \ \mathcal{C}_i\big(x^{\mathrm{ego}}(t), x_i(t)\big) = \mathrm{true}\right\}.$$

Define the distance traveled by the ego *along its planned trajectory* until that conflict:

$$D_i^\star(t_0) = \int_{t_0}^{t_i^\star} \|\dot{x}^{\mathrm{ego}}(t)\| \ dt.$$

Define the scene distance-to-first-conflict as the minimum over objects:

$$D^\star(t_0) = \min_{i \in \{1,\ldots,N\}} D_i^\star(t_0).$$

$D^\star(t_0)$ is the arc-length along the planned ego motion until the earliest predicted safety conflict, not a Euclidean range.

TABLE I
NOTATION SUMMARY (UNITS IN PARENTHESES).

Symbol	Meaning	Symbol	Meaning
t_0 (s)	current time at which DaVoS is evaluated	o (s)	end-to-end sense–decide–act overhead at t_0
t (s)	time variable for plan/predictions	t_{obs} (s)	observation timestamp used to measure overhead
u (s)	integration variable over overhead window	t_{eff} (s)	earliest time updated decision affects motion
T^{plan} (s)	planning/prediction horizon length	V	variability descriptor selecting conservative policy at t_0
Δt (s)	discretization step for conflict search	$\mathcal{Q}_V^X$	conservative transform for $X \in \{D, S, a, o\}$ under V
t_k (s)	sample times $t_k = t_0 + k\Delta t$	$D_V^{\downarrow}$ (m)	conservative lower value of $D^\star$
k, K	sample index and final index ($k = 0, \dots, K$)	$S_V^{\uparrow}$ (m/s)	conservative upper value of S
i, N	object index and number of predicted objects	$a_V^{\uparrow}$ (m/s^2)	conservative upper value of a
$^{\text{ego}}$	superscript denoting the ego vehicle	$o_V^{\uparrow}$ (s)	conservative upper value of o
$x^{\text{ego}}(t)$ (m)	planned ego trajectory (time-parameterized)	p	quantile/percentile level used in conditional quantiles and margin selection
$\dot{x}^{\text{ego}}(t)$ (m/s)	time derivative of planned ego trajectory	$q_p(X \mid V)$	p-th conditional quantile used to choose margins
$x_i(t)$ (m)	predicted trajectory of object i	$\text{median}(X \mid V)$	conditional median used in margin definitions
$v^{\text{ego}}(t_0)$ (m/s)	estimated ego velocity vector at t_0	$k_X(V)$ (units of X)	calibrated margin used in $\mathcal{Q}_V^X$
$a^{\text{ego}}(t_0)$ (m/s^2)	estimated ego acceleration vector at t_0	ΔD^{commit} (m)	committed distance over overhead window (constant-accel approx.)
$\|\cdot\|$, $\langle\cdot,\cdot\rangle$	Euclidean norm and inner product	u^{stop} (s)	time-to-stop used when $a_V^{\uparrow} < 0$
S (m/s)	ego speed magnitude at t_0, $S = \|v^{\text{ego}}(t_0)\|$	M (m)	slack $M = D_V^{\downarrow} - \Delta D^{\text{commit}}$
a (m/s^2)	signed longitudinal acceleration at t_0	d_0 (m)	distance scale in score calibration
true, false	boolean truth values used by $\mathcal{C}_i$	$\sigma(\cdot)$	monotone squashing function $\mathbb{R} \to (0,1)$ (logistic used here)
$\mathcal{C}_i(\cdot)$	conflict predicate; returns true or false	Score^{DaVoS}	DaVoS score at t_0, $= 100\,\sigma(M/d_0) \in (0, 100)$
$t_i^\star$ (s)	earliest predicted conflict time for object i (horizon-censored)	∞	used to denote "infinite" TTC when the ego is not closing
$D_i^\star$ (m)	arc-length along x^{ego} from t_0 to $t_i^\star$	TTC (s)	lane-local TTC proxy at t_0; ∞ when not closing
$D^\star$ (m)	distance-to-first-conflict at t_0, $D^\star = \min_i D_i^\star$	Score^{TTC}	TTC-based baseline score used in Figure 3
$T^\star$ (s)	implied time-to-conflict, $T^\star = D^\star/S$	$g(t)$ (m)	longitudinal headway gap to the current in-lane leader

If the plan and predictions are only defined on $[t_0,\, t_0 + T^{plan}]$, DaVoS defines the conflict search over this finite horizon. If no conflict occurs within the horizon for object i, we set $t_i^\star = t_0 + T^{plan}$ when computing $D_i^\star$ and interpret $D_i^\star$ as a horizon-censored distance. Consequently, a large value of $D^\star$ (and therefore a high score) should be interpreted as "no predicted conflict within the planning horizon," not as an unconditional safety guarantee beyond that horizon.

E. Current speed (S) and current acceleration/deceleration (a)

Let $v^{\text{ego}}(t_0)$ and $a^{\text{ego}}(t_0)$ denote the estimated velocity and acceleration vectors of the ego state. Define the current ego speed

$$S(t_0) = \|v^{\text{ego}}(t_0)\|\,.$$

Define the current ego longitudinal acceleration along the instantaneous direction of motion as a signed quantity

$$a(t_0) = \begin{cases} \frac{\langle v^{\text{ego}}(t_0),\, a^{\text{ego}}(t_0)\rangle}{\|v^{\text{ego}}(t_0)\|}, & S(t_0) > 0, \\ 0, & S(t_0) = 0, \end{cases}$$

with $a(t_0) > 0$ indicating speeding up and $a(t_0) < 0$ indicating braking/deceleration. The definition uses the acceleration estimate directly and avoids differentiating a noisy speed signal.

F. End-to-end overhead (o)

Let $o(t_0)$ denote the end-to-end sense–decide–act overhead in seconds: the delay between the world being observed and a materially different vehicle motion being realized. o aggregates sensor sampling delay, compute/queueing delay, middleware/communication latency, scheduling jitter, control computation delay, and actuation response delay. While the system is within this overhead window, the vehicle continues to move according to its current motion state; thus o directly translates into distance committed before a new decision can take effect.

G. Measured variability (V) and conservative derating

Let $V(t_0)$ summarize measured variability relevant to $D^\star$, S, a, and o under operational conditions. Operationally, $V(t_0)$ can be a discrete condition bucket or a scalar index; its role is to select which conservative policy is applied at time t_0.

DaVoS uses V to define conservative versions of the instantaneous quantities through standardized conservative transforms:

$$D_V^{\downarrow}(t_0) = \mathcal{Q}_V^D\big(D^\star(t_0)\big), \qquad S_V^{\uparrow}(t_0) = \mathcal{Q}_V^S\big(S(t_0)\big),$$
$$a_V^{\uparrow}(t_0) = \mathcal{Q}_V^a\big(a(t_0)\big), \qquad o_V^{\uparrow}(t_0) = \mathcal{Q}_V^o\big(o(t_0)\big).$$

The arrows indicate the conservative direction: $D_V^{\downarrow}$ is a conservative *lower* value of distance-to-conflict (smaller is worse). $S_V^{\uparrow}$ is a conservative *upper* value of speed (larger is worse). $a_V^{\uparrow}$ is a conservative *upper* value of acceleration (more positive, i.e. less braking or more speeding-up, is worse). $o_V^{\uparrow}$ is a conservative *upper* value of overhead (larger is worse).

For reproducibility and comparability, $\mathcal{Q}_V^X : \mathbb{R} \to \mathbb{R}$ must be a deterministic mapping applied to the instantaneous scalar $X(t_0)$ (i.e., it is not "take a quantile of a scalar at runtime"). A simple and reproducible choice is a margin-based transform with nonnegative margins $k_X(V)$ calibrated offline from a disclosed dataset:

$$\begin{aligned} \mathcal{Q}_V^D(d) &= \max\{0,\, d - k_D(V)\}, \\ \mathcal{Q}_V^S(s) &= s + k_S(V), \\ \mathcal{Q}_V^a(\alpha) &= \alpha + k_a(V), \\ \mathcal{Q}_V^o(o) &= o + k_o(V), \end{aligned} \tag{1}$$

where $k_X(V) \geq 0$ is fixed for a given variability bucket V. Quantiles enter through how margins are chosen: for example, one may set $k_o(V) = q_p(o \mid V) - \text{median}(o \mid V)$ for a disclosed p and calibration procedure, and analogously for S and a; for $D^\star$, one may set $k_D(V) = \text{median}(D^\star \mid V) - q_{1-p}(D^\star \mid V)$. Here $q_p(X \mid V)$ denotes the p-th quantile of X conditioned on V. The definition of V, the calibration dataset, and the choice of p (or an equivalent margin-selection rule) are part of the DaVoS specification and must be disclosed for comparisons.

Because $D^\star$, S, a, and o can be correlated, applying conservative transforms marginally does not, in general, guarantee a joint-tail bound on the slack. If stricter conservatism is required, one may calibrate conservatism directly on the slack M conditioned on V.

The conservative policy (including quantile levels or equivalent margin rules) is part of the metric specification and must be disclosed.

H. Committed distance over the overhead window

Over the overhead window of length $o_V^{\uparrow}(t_0)$, DaVoS uses the current motion state $(S_V^{\uparrow}, a_V^{\uparrow})$ to approximate the distance traveled before a new decision can take effect.

Define the committed distance as the time-integral of the (nonnegative) speed under constant acceleration:

$$\Delta D^{\text{commit}}(t_0) = \int_0^{o_V^{\uparrow}(t_0)} \max\Big\{0,\; S_V^{\uparrow}(t_0) + a_V^{\uparrow}(t_0)u\Big\}\, du. \tag{2}$$

Equation (2) is a standardized approximation. It is conservative when realized speed and longitudinal acceleration are bounded above by $(S_V^{\uparrow}, a_V^{\uparrow})$ over the overhead window; otherwise, it should be interpreted as a standardized estimate.

When $a_V^{\uparrow}(t_0) \geq 0$ (speeding up or coasting), speed will not decrease to zero over the interval and the integral reduces to

$$\Delta D^{\text{commit}}(t_0) = S_V^{\uparrow}(t_0)\, o_V^{\uparrow}(t_0) + \frac{1}{2} a_V^{\uparrow}(t_0)\big(o_V^{\uparrow}(t_0)\big)^2. \tag{3}$$

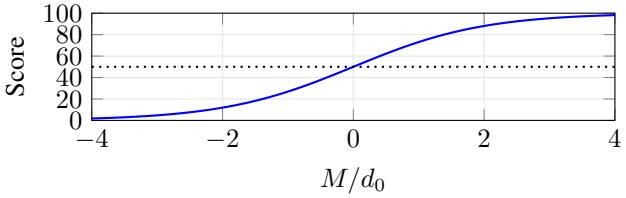

Fig. 2. Sigmoid calibration: $\text{Score}^{DaVoS} = 100\,\sigma(M/d_0)$.

When $a_V^{\uparrow}(t_0) < 0$ (braking), define the time-to-stop under the constant-acceleration approximation:

$$u^{\text{stop}}(t_0) = \min\left\{ o_V^{\uparrow}(t_0),\; \frac{S_V^{\uparrow}(t_0)}{-a_V^{\uparrow}(t_0)} \right\}.$$

Then the committed distance reduces to

$$\Delta D^{\text{commit}}(t_0) = S_V^{\uparrow}(t_0)\, u^{\text{stop}}(t_0) + \frac{1}{2} a_V^{\uparrow}(t_0)\big(u^{\text{stop}}(t_0)\big)^2. \tag{4}$$

$\Delta D^{\text{commit}}(t_0)$ is the distance the ego is expected to travel before the system can materially change behavior, given its current speed and acceleration/deceleration and given conservative overhead. It increases with higher speed, higher (more positive) acceleration, and higher overhead. It decreases when the current state reflects braking ($a < 0$).

I. Safety score and calibration

DaVoS defines an instantaneous safety-confidence score by comparing the conservative distance-to-first-conflict with the conservative committed distance. Define

$$\text{Score}^{DaVoS}(t_0) = 100 \cdot \sigma\left(\frac{D_V^{\downarrow}(t_0) - \Delta D^{\text{commit}}(t_0)}{d_0} \right). \tag{5}$$

where $d_0 > 0$ is a distance scale and $\sigma(\cdot)$ is any monotone squashing function mapping $\mathbb{R} \to (0, 1)$. A common choice is the logistic sigmoid

$$\sigma(z) = \frac{1}{1 + e^{-z}}.$$

With this choice, $\text{Score}^{DaVoS}(t_0) \in (0, 100)$ and approaches the endpoints asymptotically; the distance scale d_0 controls how rapidly the score transitions with slack (Figure 2).

If $D_V^{\downarrow}(t_0) \gg \Delta D^{\text{commit}}(t_0)$, the slack is positive and the score approaches 100; if $D_V^{\downarrow}(t_0) < \Delta D^{\text{commit}}(t_0)$, the slack is negative and the score approaches 0.

Figure 2 illustrates this mapping for the logistic choice of σ. The horizontal axis is the *normalized slack* M/d_0, so $M = 0$ maps to $\text{Score}^{DaVoS} = 50$ (the conservative boundary), and one unit on the horizontal axis corresponds to d_0 meters of slack. Thus d_0 is an explicit interpretability knob: smaller d_0 yields a sharper transition near $M = 0$, while larger d_0 produces a smoother score curve for the same changes in margin. For reference, under the logistic curve in Figure 2, $M \approx 2.2\, d_0$ maps to $\text{Score}^{DaVoS} \approx 90$ and $M \approx -2.2\, d_0$ maps to $\text{Score}^{DaVoS} \approx 10$.

In the score, the distance-to-conflict term appears as $D_V^{\downarrow}(t_0)$, derived from ego plan and object predictions; S and a appear inside $\Delta D^{\mathrm{commit}}(t_0)$ as the current motion state; o appears as the overhead window length $o_V^{\uparrow}(t_0)$; and V determines how conservative each term becomes by selecting the operators $\mathcal{Q}_V^D, \mathcal{Q}_V^S, \mathcal{Q}_V^a, \mathcal{Q}_V^o$.

The score is monotone in the intended directions: it increases with larger conservative distance-to-conflict $D_V^{\downarrow}$ and decreases with larger conservative overhead $o_V^{\uparrow}$, speed $S_V^{\uparrow}$, and acceleration $a_V^{\uparrow}$, because each of these increases the conservative committed distance or decreases available slack. This monotonicity is a design goal: it ensures that improvements in timing tails, speed policy, or prediction/planning that postpone conflicts are reflected as score increases under a fixed conservative policy.

Define the distance slack

$$M(t_0) = D_V^{\downarrow}(t_0) - \Delta D^{\mathrm{commit}}(t_0).$$

Then $\mathrm{Score}^{DaVoS}(t_0) = 100\,\sigma\big(M(t_0)/d_0\big)$, so d_0 (meters) sets how quickly the score transitions with slack. In the special case of identity conservative transforms, constant speed, and $a(t_0) = 0$, the committed distance reduces to $S(t_0)o(t_0)$ and therefore $M(t_0) = D^\star(t_0) - S(t_0)o(t_0) = S(t_0)\big(T^\star(t_0) - o(t_0)\big)$ where $T^\star(t_0) = D^\star(t_0)/S(t_0)$ is the implied time-to-conflict along the plan. In this sense, DaVoS can be viewed as a latency-aware generalization of time-to-conflict and related margin-based proxies, with V controlling conservatism and with acceleration entering through the committed-distance term. These proxies and their empirical use in safety analysis are widely studied in traffic safety literature [2], [3]. For the logistic sigmoid, $\mathrm{Score}^{DaVoS} = 50$ exactly when $M = 0$, and if one desires $\mathrm{Score}^{DaVoS} = 100p$ at slack $M = M_p$, then

$$d_0 = \frac{M_p}{\ln\!\left(\frac{p}{1-p}\right)}.$$

For example, $\mathrm{Score}^{DaVoS} = 90$ at $M = M_{90}$ implies $d_0 = M_{90}/\ln(9)$.

IV. Interpretation and Examples

The model is intentionally compact so that changes in system design map to changes in a small number of measured quantities. Holding the scenario set, conflict predicate, and conservative policy fixed, reductions in end-to-end timing tails decrease $o_V^{\uparrow}$ and therefore decrease committed distance, increasing the score. Less aggressive acceleration near conflicts decreases $a_V^{\uparrow}$, and speed-policy changes affect $S_V^{\uparrow}$ directly. Improvements in planning and prediction that postpone or eliminate predicted conflicts increase $D_V^{\downarrow}$. Finally, reduced variability tightens conservative transforms, increasing $D_V^{\downarrow}$ while decreasing $S_V^{\uparrow}$, $a_V^{\uparrow}$, and $o_V^{\uparrow}$ under a fixed specification.

A. Highway car-following scenario (30 s)

We consider a straight-line highway car-following scene. The ego gradually closes the gap to the in-lane lead vehicle and then maintains a minimum longitudinal headway of 50 m. Let $g(t)$ denote the longitudinal gap along the lane centerline from ego to the current in-lane leader.

For this example, $\mathcal{C}_i$ encodes a longitudinal headway boundary: a conflict is predicted when the in-lane separation falls below 50 m. We treat $V(t_0)$ as nominal so the conservative transforms reduce to identities. We use $d_0 = 100$ m, $T^{plan} = 8$ s, and $\Delta t = 0.5$ s, and we plot curves for $o \in \{0.2, 0.6, 1.0\}$ s.

For comparison, we plot a latency-unaware TTC-based baseline score that ignores acceleration: $\mathrm{Score}^{TTC}(t_0) = 100 \cdot \sigma\Big(\frac{S(t_0)\min\{\mathrm{TTC}(t_0), T^{plan}\}}{d_0}\Big)$, where $\mathrm{TTC}(t_0)$ is the time until the headway boundary would be violated under constant relative speed (treated as ∞ when the ego is not closing). TTC uses only the current in-lane leader at t_0 (no anticipated lane changes).

Figure 3**(a)** sketches the setting and the constraint $g(t) \geq 50$ m. In Figure 3**(b)**, the ego closes from an initial 110 m gap (ego 30 m/s, lead 27 m/s), then brakes from $t = 18$ s to $t = 22$ s at $a = -0.75$ m/s^2 to match speed and maintain 50 m. During the approach, $D^\star$ enters the horizon and DaVoS decreases; larger o shifts DaVoS downward. Once closing stops, the score recovers because the conflict is no longer predicted within the finite horizon.

B. Adjacent-lane cut-in scenario (30 s)

We next consider a straight-line highway cut-in, highlighting the plan- and prediction-conditioned nature of $D^\star$: DaVoS can decrease *before* the cut-in is realized when the predictor forecasts the merge within the horizon, whereas a lane-local TTC proxy does not anticipate the event.

In Figure 3**(c)**, the cut-in becomes the in-lane leader at $t = 10$ s with a 65 m headway at 24 m/s. The ego travels at 30 m/s until $t = 10.5$ s and then brakes at $a = -3$ m/s^2. Under the pre-brake plan, the 50 m boundary would be reached at approximately $t = 12.5$ s, bringing the predicted conflict inside the horizon and lowering DaVoS even before $t = 10$ s. After braking begins, the plan-conditioned $D^\star$ returns to the horizon, while the TTC baseline (ignoring acceleration) remains pessimistic during the transient.

C. From instantaneous score to scenario-level curve

DaVoS can be evaluated over time to produce a scenario-level score curve; aggregations (e.g., minimum score, time below a disclosed threshold, or exposure-normalized time-below-threshold per mile) provide compact summaries for monitoring and post-incident analysis.

V. Measurement and Implementation

$D^\star(t_0)$ is computed from the ego plan and object predictions by evaluating $\mathcal{C}_i$ over a disclosed horizon T^{plan} with sampling resolution Δt. Define $t_k = t_0 + k\Delta t$ for $k = 0, \ldots, K$ with $t_K = t_0 + T^{plan}$. For each object i, set $k_i^\star = \min\{k : \mathcal{C}_i(x^{\mathrm{ego}}(t_k), x_i(t_k)) = \mathrm{true}\}$ (or K if none) and compute $D^\star(t_0) = \min_i \sum_{k=0}^{k_i^\star - 1} \|\dot{x}^{\mathrm{ego}}(t_k)\|\,\Delta t$. Reports should disclose $(T^{plan}, \Delta t)$ and whether $D^\star$ is horizon-censored.

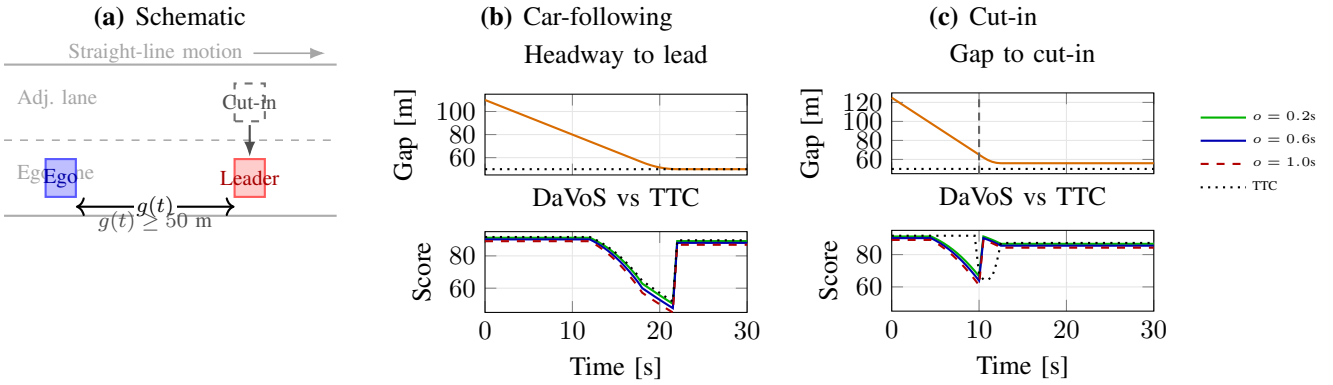

Fig. 3. Illustrative examples (synthetic traces). **Car-following:** conflict boundary is a 50 m minimum longitudinal headway to the lead vehicle in the ego lane. **Cut-in:** a neighboring vehicle merges into the ego lane at $t = 10$ s (vertical line); DaVoS decreases in advance when the cut-in is predicted within the horizon, while the TTC baseline (defined using only the current in-lane leader at t_0, i.e., no anticipated lane changes) does not anticipate the event. Under this TTC definition, the cut-in vehicle is out-of-lane for $t < 10$ s and TTC is treated as ∞ until the lane change completes. In both cases, larger end-to-end overhead o shifts DaVoS downward. Top plots show gap and the 50 m boundary; bottom plots show score curves (including the TTC baseline).

Because $\mathcal{C}_i$ is defined on continuous trajectories, discrete sampling can miss between-sample conflicts. Implementations should use conservative bracketing over $[t_k, t_{k+1}]$ or select Δt via a disclosed stability test (e.g., halve Δt until $D^\star$ changes by less than a tolerance).

$S(t_0)$ is computed as the estimated velocity magnitude, and $a(t_0)$ as the longitudinal projection $\langle v^{\text{ego}}(t_0), a^{\text{ego}}(t_0)\rangle / \|v^{\text{ego}}(t_0)\|$ (or 0 at zero speed). $o(t_0)$ is obtained from end-to-end instrumentation; for cross-platform comparability, define o as the latency from an observation timestamp to the earliest time when an updated decision can begin to affect realized motion, under a disclosed clock-synchronization policy. Latency and input-delay effects of this form are widely studied in real-time control and networked autonomy [14], [15].

Operationally, log two timestamps per decision cycle: an observation timestamp t_{obs} and an effect timestamp t_{eff} defined as the earliest time at which the updated decision is observable in actuation or vehicle response, using a disclosed threshold for "materially different." Define $o = t_{\text{eff}} - t_{\text{obs}}$. If only command-publication timestamps are available, define an approximate o relative to the first control-command publication time and report the omission of downstream delays.

$V(t_0)$ is a disclosed variability descriptor (e.g., a discrete condition bucket over compute platform, thermal regime, friction class, and payload class) that selects the conservative transforms $\mathcal{Q}_V^{(\cdot)}$ used to produce $D_V^{\downarrow}$, $S_V^{\uparrow}$, $a_V^{\uparrow}$, and $o_V^{\uparrow}$. For comparison, publish the mapping from measured conditions to V, the transform definitions (e.g., which quantiles are used in each bucket), and the calibration dataset/procedure used to estimate them.

For reporting, disclose the metric specification (scenario set, $\mathcal{C}_i$, $(T^{plan}, \Delta t)$, o definition, conservative transforms, and score calibration) and report both instantaneous curves and at least one scenario-level aggregation. Unless otherwise specified, we recommend reporting the minimum score over time together with the duration spent below a disclosed threshold.

VI. Limitations and Extensions

DaVoS is only as meaningful as the plan, predictions, and conflict predicate used to compute $D^\star$. If the planner or predictor omits relevant agents or behaviors, or if $\mathcal{C}_i$ does not encode the intended safety boundary, then $D^\star$ (and therefore the score) can be misleading. For this reason, comparisons must treat $\mathcal{C}_i$ and the scenario set as part of the metric specification rather than as incidental implementation details.

Because the model uses a finite planning/prediction horizon, "no predicted conflict" should be interpreted as "no predicted conflict within the horizon." In high-speed scenes or in scenes with long latency, the chosen horizon can materially change the score; reporting should therefore disclose the horizon and any truncation convention.

The committed-distance approximation in (2) treats speed evolution over the overhead window using a constant-acceleration model derived from the current state. This choice is intentional: it yields a standardized instantaneous baseline that is always well-defined from the five DaVoS quantities.

The base definition measures conservative headroom to initiate a materially different action under the current plan and predictions; avoidability beyond the overhead window requires additional modeling and is outside the scope of the compact instantaneous score.

VII. Conclusion

DaVoS defines an instantaneous safety-confidence score by comparing conservative distance-to-first-predicted-conflict along the ego plan with conservative committed distance over the end-to-end overhead window. By expressing confidence in physically interpretable quantities and measured timing overhead, DaVoS provides a compact reference for diagnostics and comparative evaluation under a disclosed metric specification.

VIII. Acknowledgment

This work is partially supported by the US National Science Foundation NSF-2311087.

REFERENCES

[1] N. Kalra and S. M. Paddock, "Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?" RAND Corporation, Tech. Rep. RR-1478-RC, 2016. [Online]. Available: https://www.rand.org/pubs/research_reports/RR1478.html

[2] J. C. Hayward, "Near miss determination through use of a scale of danger," *Highway Research Record*, no. 384, 1972.

[3] K. Vogel, "A comparison of headway and time to collision as safety indicators," *Accident Analysis & Prevention*, vol. 35, no. 3, pp. 427–433, 2003.

[4] P. Koopman and M. Wagner, "Autonomous vehicle safety: An interdisciplinary challenge," *IEEE Intelligent Transportation Systems Magazine*, vol. 9, no. 1, pp. 8–11, 2017.

[5] A. Boggs, R. Arvin, and A. J. Khattak, "Exploring the who, what, when, where, and why of automated vehicle disengagements," *Accident Analysis & Prevention*, vol. 136, p. 105406, 2020.

[6] S. Dixit, S. Fallah, U. Montanaro, M. Dianati, A. Stevens, D. Oxtoby, and A. Mouzakitis, "Trajectory planning and tracking for autonomous vehicles: A disengagement perspective," *PLOS ONE*, vol. 11, no. 12, p. e0168054, 2016.

[7] California Department of Motor Vehicles, "Disengagement reports," Web page, 2026, accessed 2026-02-22. [Online]. Available: https://www.dmv.ca.gov/portal/vehicle-industry-services/autonomous-vehicles/disengagement-reports/

[8] *ISO 26262-1:2018 Road vehicles — Functional safety — Part 1: Vocabulary*, International Organization for Standardization (ISO) Std., 2018.

[9] *ISO 21448:2022 Road vehicles — Safety of the intended functionality*, International Organization for Standardization (ISO) Std., 2022.

[10] *ANSI/UL 4600: Standard for Safety for the Evaluation of Autonomous Products*, UL Standards & Engagement Std., 2020.

[11] P. G. Gipps, "A behavioural car-following model for computer simulation," *Transportation Research Part B: Methodological*, vol. 15, no. 2, pp. 105–111, 1981.

[12] S. Shalev-Shwartz, S. Shammah, and A. Shashua, "On a formal model of safe and scalable self-driving cars," arXiv preprint arXiv:1708.06374, 2017.

[13] P. Koopman, B. Osyk, and J. Weast, "Autonomous vehicles meet the physical world: Rss, variability, uncertainty, and proving safety," arXiv preprint arXiv:1911.01207, 2019.

[14] I. Batković, M. Zanon, M. Ali, and P. Falcone, "Real-time constrained trajectory planning and vehicle control for advanced driver assistance systems with input delays," in *2019 18th European Control Conference (ECC)*, 2019, pp. 256–262.

[15] M. Narasimhan, A. Gahlawat, H. Chen, S. Zhai, M. Pavone, and S. B. Das, "Safe networked robotics with probabilistic verification: Stochastic delay and packet dropout," arXiv preprint arXiv:2302.09182, 2023.