Synthesizing Guide Programs for Sound, Effective Deep Amortized Inference

JIANLIN LI, University of Waterloo, Canada
LENI VEN, University of Waterloo, Canada
PENGYUAN SHI, University of Waterloo, Canada
YIZHOU ZHANG, University of Waterloo, Canada

In probabilistic programming languages (PPLs), a critical step in optimization-based inference methods is constructing, for a given model program, a trainable guide program. Soundness and effectiveness of inference rely on constructing good guides, but the expressive power of a universal PPL poses challenges. This paper introduces an approach to automatically generating guides for deep amortized inference in a universal PPL. Guides are generated using a type-directed translation per a novel behavioral type system. Guide generation extracts and exploits independence structures using a syntactic approach to conditional independence, with a semantic account left to further work. Despite the control-flow expressiveness allowed by the universal PPL, generated guides are guaranteed to satisfy a critical soundness condition and, moreover, consistently improve training and inference over state-of-the-art baselines for a suite of benchmarks.

This technical report accompanies Li et al. [2023].

1 INTRODUCTION

A Bayesian probabilistic model denotes a distribution \( p(Z, X) \), where random variables \( Z \) are called latent and \( X \) observed. Bayesian inference is aimed at calculating \( p(Z|X = x) \), the posterior distribution of latents \( Z \) given observed data \( x \).

Probabilistic programming languages (PPLs) are powerful tools for modeling Bayesian-inference problems. Universal PPLs, in particular, offer linguistic features including stochastic branching and general recursion, making it possible to express mixture models, probabilistic context-free grammars, kernel induction models [Le et al. 2019; Saad et al. 2019], and so forth.

PPLs also provide inference methods for solving inference problems. Optimization-based methods, such as variational inference [Jordan et al. 1999], are taking hold in some PPLs empowered by deep neural networks [Bingham et al. 2019; Tran et al. 2018]. They consist in approximating the true posterior \( p(Z|X = x) \) of a model program using a neural guide program \( q_\phi(Z) \) where \( \phi \) stands for neural-network parameters: they optimize an objective function via gradient descent, searching for \( \phi \) that makes \( q_\phi(Z) \) close to \( p(Z|X = x) \). Notice that guide programs must be free of observed random variables.

Deep amortized inference [Paige and Wood 2016; Ritchie et al. 2016; Le et al. 2017; van de Meent et al. 2021, §8] is based on optimization and profits from learning amortized guides. Amortization is an ingrained idea in machine learning literature [Hinton et al. 1995; Kingma and Welling 2014]. Rather than training neural networks every time a new observation is given, amortized inference trains ahead of time a guide program \( q_\phi(Z; x) \) that takes an observation \( x \) as input. Then, at run time, inferring \( p(Z|X = x_0) \) for actual observations \( x_0 \) is cheap using \( q_\phi(Z; x_0) \), amortizing the upfront cost of training \( \phi \). Optimization and amortization help inference scale for probabilistic programs that need to be applied repeatedly to different observations.

Constructing good guides is critical. Ahead-of-time training and run-time inference are (1) sound only when model–guide pairs satisfy absolute continuity and (2) most effective when guides are faithful yet parsimonious in terms of the conditional dependences they encode:

1. Absolute continuity, informally speaking, is a soundness condition requiring two distributions to have compatible supports. A guide with an incompatible support causes inference to crash or produce incorrect results [Lee et al. 2019].
Faithful guides do not contain conditional independences not found in models—missing correlations make guides unable to express true posteriors even if they had unbounded neural-network capacities (i.e., number of trainable parameters). Parsimonious guides do not encode more conditional dependences than necessary—excessive correlations introduce computational burdens as ahead-of-time training has to unlearn them to uncover true posteriors [Webb et al. 2018].

Mainstream PPLs currently have limited support for emitting guide programs. In Pyro, for example, the autoguide [2022] library works only for non-universal probabilistic programs, does not guarantee absolute continuity, and ignores faithfulness altogether. Creating good guide programs remains a demanding, error-prone task.

This paper introduces an approach to automatically generating guide programs for a universal PPL. Generated guides enjoy strong soundness guarantees and consistently improve training and inference over state-of-the-art approaches, despite the expressiveness allowed by the universal PPL.

We first review relevant background in Section 2. We then identify technical challenges and our core contributions in Section 3.
Addresses are unique names identifying RVs in a trace. Consider the recursive function pecfg (written in Pyro) in Figure 2a, which models a probabilistic context-free grammar. A possible trace of this program is
\[
["_A": .7, "_D1_A": .3, "_D1_C": 2, "_D2_A": .1, "_D2_C": 9],
\]
which will result in the expression \( \text{Add} (\text{Const}(2), \text{Const}(9)) \) in the context-free language.

Deep amortized inference. The ahead-of-time training stage of amortized inference searches for \( \phi \) that makes \( q_\phi (Z; x) \) similar to \( p(Z|X = x) \) on average for most \( x \). More precisely, it solves the following optimization problem:
\[
\arg\min_\phi \mathbb{E}_{x \sim p(X)} \left[ \text{KL}(p(Z|x) \| q_\phi (Z; x)) \right],
\]
where \( p(Z|x) \) abbreviates \( p(Z|X = x) \), and the Kullback–Leibler divergence quantifies how similar two distributions are. Objective (2.2) can be simplified to
\[
\arg\min_\phi \mathbb{E}_{z, x \sim p(Z, X)} \left[ - \log q_\phi (z; x) \right],
\]
which is convenient: training data \( z, x \sim p(Z, X) \) can be generated simply by running the model program \( p(Z, X) \) repeatedly. We note that a variant of amortized inference [Ritchie et al. 2016] optimizes a different objective, known as the evidence lower bound (ELBo):
\[
\arg\max_\phi \prod_{x_i} \mathbb{E}_{z \sim q_\phi (Z; x_i)} \left[ \log p(z, x_i) - \log q_\phi (z; x_i) \right].
\]
Objective (2.4) requires a training dataset \( \{x_1, \ldots, x_n\} \) be provided, as indicated by the summation over \( x_i \) in (2.4). In this paper, we focus on the first type of amortized inference that uses objective (2.3), but our technical contributions can in principle be applied to the ELBo objective (2.4).

The run-time inference stage performs importance sampling (IS), per the equation
\[
p(Z|x_0) = p(Z, x_0) \left/ \mathbb{E}_{z \sim q_\phi (Z|x_0)} \left[ \frac{p(z, x_0)}{q_\phi (z; x_0)} \right] \right\}.
\]
Given observation \( x_0 \), IS approximates \( p(Z|x_0) \) by sampling \( z \) from the trained proposal distribution \( q_\phi (Z; x_0) \) and weighting samples \( z \) by their importance ratio \( p(z, x_0)/q_\phi (z; x_0) \).

3 PROBLEMS AND CONTRIBUTIONS

3.1 Absolute Continuity (AC)

Crucially, soundness of both training and inference relies on AC. First, per formula (2.5), IS is unbiased only when \( p(Z, x_0) \) is absolutely continuous with respect to \( q_\phi (Z; x_0) \), written as \( p(Z, x_0) \ll q_\phi (Z; x_0) \). That is, in a trace-based PPL, for samples from \( q_\phi (Z; x_0) \) to asymptotically approximate the posterior, \( q_\phi (Z; x_0) \) is required to have non-null probability on any measurable set of traces on which \( p(Z, x_0) \) has non-null probability. Second, objective (2.3) is defined only when \( p(Z, x) \ll q_\phi (Z; x) \), lest logarithm-of-zero errors.\(^1\) Furthermore, the ELBo objective (2.4) is defined only when AC holds in the other direction: \( q_\phi (Z; x) \ll p(Z, x) \). Indeed, AC is a pervasive soundness condition required by many inference methods (e.g., MCMC).

It is simple to verify AC when probabilistic programs can be described as BNs: it suffices to check that two programs contain the same set of RVs. For example, in Figure 1, model is absolutely continuous with respect to both guides, even though the guides reverse the order \( a \) and \( b \) are sampled. All three programs have the same support made up of two RVs: \( a \) over \( \mathbb{R} \) and \( b \) over \( \mathbb{R}_+ \).

Control-flow expressiveness of universal PPLs poses challenges to a formal argument, however: branching and recursion allow a stochastic set of RVs to be sampled every time a program is run.

\(^1\) Gradient descent further requires differentiability, which is studied in other work (e.g., Mak et al. [2021]).
\[ \text{def } \text{model (obs)} = \]
\[ a = \text{sample(Normal}(0; 2)) \]
\[ b = \text{sample(Gamma}(8; 2)) \]
\[ \text{observe(Normal}(a+b, 0.05); \text{obs}) \]

\[ \text{def } \text{guide1 (obs)} = \]
\[ x_1, x_2 = \text{NNb}(\text{obs}) \]
\[ b = \text{sample(Gamma}(x_1; x_2)) \]
\[ y_1, y_2 = \text{NNa}(\text{obs}) \]
\[ a = \text{sample(Normal}(y_1; y_2)) \]

\[ \text{def } \text{guide2 (obs)} = \]
\[ x_1, x_2 = \text{NNb}(\text{obs}) \]
\[ b = \text{sample(Gamma}(x_1; x_2)) \]
\[ y_1, y_2 = \text{NNa}(\text{obs}, b) \]
\[ a = \text{sample(Normal}(y_1; y_2)) \]

Figure 1. Soundness: model is (trivially) absolutely continuous with both guide1 and guide2, despite that the guides reorder \(a\) and \(b\). Faithfulness: guide2 faithfully captures conditional dependence between \(a\) and \(b\) given \(\text{obs}\), whereas guide1 considers \(a\) and \(b\) independent given \(\text{obs}\). The payoff of being faithful is that guide2 leads to higher-quality posterior samples than guide1.

Consider the guide in Figure 2b, manually created for the PCFG model in Figure 2a. It might appear that the guide is compatible with the model: half the time it samples a leaf node of the syntax tree, and the other half the time it recursively generates subtrees. But it has two sources of unsoundness:

- Its branching condition \(a > .5\), different from the model’s, causes the guide to execute different branches—and thus sample different RVs—than the model. The support mismatch causes accesses to unavailable addresses, crashing a typical inference engine. Notice, however, that reordering of the subtrees by the guide does not change the distribution’s support and is thus not unsoundness.

- Second, RVs at addresses \(L + "C"\) have a smaller support \((R)\) than they do in the model \((R)\). The mismatch can lead to logarithm-of-zero errors, NaNs, or biased estimates.

The run-time errors are frustrating when they happen deep into the gradient-descent loop. Even worse is when programs return normally but with incorrect results.

State-of-the-art approaches. The challenge has led to recent work on type systems for ensuring AC statically [Lew et al. 2019; Wang et al. 2021]: a model and a guide are guaranteed to satisfy AC if they have compatible types. Unfortunately, these systems are not geared to automatic guide generation, and more critically, they restrict expressible models or guides. Lew et al.’s type system supports only limited forms of branching and loops, and it disallows general recursion. Wang et al. address specifically these limitations, but their coroutine-based semantics requires a guide to perform computations in exactly the same order as its model does (thus failing to type-check trivially sound guides such as guide1 and guide2 in Figure 1).

These restrictions have implications to modeling power and to performance. Without general recursion, it is difficult to express useful models including probabilistic context-free grammars [Manning and Schütze 1999]. Without the possibility to reorder RVs, faithful guides may require denser dependence structures than otherwise avoidable (see Figure 16, for example), which can reduce training performance as we show later.

3.2 Faithfulness and Parsimony

Faithfulness matters. Effectiveness of IS is highly dependent on having good proposals. In amortized inference, IS yields low-variance estimates if ahead-of-time training makes the proposal distribution
**Figure 2.** Implementation of a probabilistic context-free grammar (PCFG). The guide breaks AC as required by formulas (2.2) and (2.5). It is unfaithful, as it introduces conditional independences not found in the model.

$q_\phi(Z; x_0)$ similar in shape to the true posterior $p(Z|x_0)$. But an unfaithful guide could not represent the true posterior even if it had infinite neural-network capacities, because it cannot encode some conditional dependences present in the model.

In Figure 1, guide2 faithfully expresses that $a$ and $b$ are conditionally dependent given $obs$ using the highlighted code. By contrast, guide1 considers $a$ and $b$ independent. Thus, after training neural networks $NN_a$ and $NN_b$, guide2 can approximate true posteriors much better than guide1, as shown by plots $q_2(a, b)$ and $q_1(a, b)$. In turn, at run time, IS with guide2 leads to much higher-quality samples than IS with guide1, as shown by plots $\hat{p}_2(a, b|obs)$ and $\hat{p}_1(a, b|obs)$.

Parsimony matters too. Parsimonious guides do not encode more conditional dependences than necessary. Given that training budgets and network capacities are finite, it would be wasteful to expend resources on unlearning spurious correlations that could otherwise be ruled out from the get-go. As we show later, parsimonious guides can indeed lead to better convergence.

Notice that parsimony of dependences is not in conflict with overparameterization of neural networks. Overparameterized networks have many more trainable parameters than there are training examples. It has been observed that they can make training easier while improving generalizability [Zhang et al. 2017]. A parsimonious guide can choose to overparameterize its neural networks.

**State-of-the-art approaches.** Methods for automating the design of faithful guides exist in the machine learning literature. Webb et al. [2018] construct faithful, parsimonious guides for BN models, but the approach is not applicable to universal PPLs with branching or recursion. For instance, the approach cannot handle the PCFG model in Figure 2a.

In fact, it is not entirely clear how to even manually create a faithful, parsimonious guide for the PCFG model, because of general recursion. Unsoundness aside, the guide in Figure 2b is unfaithful. It is based on the simplifying mean-field assumption that RVs are conditionally independent (given $obs$). Mean-field guides are a standard default in practice for variational inference [Zhang et al. 2019], but as our experiments confirm, they may lead to unsatisfying results for PCFG-like models that exhibit strong correlations between RVs.

To handle branching and recursion, Le et al. [2017] construct guides using advanced recurrent neural networks (RNNs) such as an LSTM, effectively treating any two random variables as correlated. Thus, RNN guides may be non-parsimonious, and so ahead-of-time training must work extra hard to uncover independences. As we also show, RNNs’ tendency to forget long-range
dependences may weaken its correlation-preserving guarantee for RVs sampled far apart. What is
needed of a guide generator is the ability to take into account structures of probabilistic programs.

3.3 Contributions

Figure 3 shows how our guide generator (called Fidelio) works. In Section 4, we design a behavioral
trace-type system capturing probabilistic effects and control flow—compatible typing implies
absolute continuity. In Section 5, we construct dependence graphs for probabilistic programs
(PDGs) and adapt the notion of active trails from BNs to PDGs—active trails suggest possible
correlation. Guide generation is then a type-guided, dependence-aware translation (Section 6).

A unique combination of features in the Fidelio semantics enables it to offer soundness guarantees
while permitting expressiveness. First, traces are tree-structured rather than lists. So a guide can
reorder computations that are subtrees of a parent, while generating the same set of traces as its
model. Second, trace types are lightweight dependent types that can track control flow through
what we call checkpoint expressions. So a model is free to use stochastic branching and general
recursion, while being well-typed.

We prove that our type system is sound: compatibly typed model–guide pairs satisfy AC (The-
orem 4.5). We prove that generated guides preserve typing (Theorem 6.1). As a result, AC is
guaranteed by construction (Theorem 6.2). We also show that our type system allows more models
and guides to be typed using a collection of programs from literature (Section 8.1).

Incorporating program dependence information in guide generation is new in itself; existing
approaches assume either independence or correlation altogether. Our use of PDGs, augmented
with hidden states and empowered by the idea of active trails, enables modularly capturing control
dependences that cannot otherwise be expressed by BNs. We note that this treatment of conditional
independence is syntactic and not proven sound; we leave a semantic account to future work.

Our evaluation of the dependence analysis focuses on assessing whether it benefits training
and inference in practice (Section 8.2). Experiments show that being dependence-aware in guide
generation pays off: Fidelio-generated guides consistently improve on state-of-the-art baselines for
a diverse set of inference problems.

4 ABSOLUTE CONTINUITY BY TYPING

In Section 4, we establish a reasoning principle (Theorem 4.5) for proving absolute continuity of
models and guides, based on typing. Section 6 then uses the reasoning principle to derive the static
guarantee of absolute continuity for automatically generated guides.

As alluded to in Section 3, our system addresses AC while permitting expressiveness, by using
tree-structured traces and checkpoint-dependent trace types. We formalize this semantics using the
Fidelio Core Calculus (FCC), which captures core aspects of the language supported by Fidelio.

4.1 Syntax of FCC

Figure 4 defines the syntax for FCC programs. Identifiers (i.e., variables and global names) are
notated in blue. Capture-free substitution is notated \( \cdot \{ / \cdot \} \): for example, \( m \{ v / a \} \) substitutes a
value \( v \) for a term variable \( a \) in command \( m \).
This section defines a trace-based, big-step operational semantics, in a style similar to prior trace-based semantics for universal PPLs (e.g., Borgström et al. [2016]). But unlike the prior semantics, traces in FCC are tree-structured rather than lists, and they can record not just sampling but also control-flow events while avoiding imposing a total ordering on these events.

Expressions evaluate to values. Judgments have form \( e \Downarrow v \). Rules in this pure fragment are standard and omitted herein. A complete semantics can be found in Appendix A.
\[
\begin{align*}
\sigma & : \text{fold } a_0 \\
\sigma_0 & : \{ a : \text{atom } .7, b : \text{inj } \{d_1 : \text{fold } \sigma_1, d_2 : \text{fold } \sigma_2\} \} \\
\sigma_1 & : \{ a : \text{atom } .3, b : \text{inj } \{c : \text{atom } 2\} \} \\
\sigma_2 & : \{ a : \text{atom } .1, b : \text{inj } \{c : \text{atom } 9\} \}
\end{align*}
\]

Figure 5. (a) PCFG model as in Figure 2a, in FCC syntax. (b) Generated guide, with the same trace type.

Figure 6. Operational semantics for the monadic fragment of FCC: terms and commands.

Terms are evaluated on events, and commands are evaluated on traces. As Figure 4 shows, events \(s\) and traces \(\sigma\) have mutually recursive syntax.

A trace \(\{a : s\}\) is an unordered map from term variables to events. Events have four forms, corresponding to the four forms of terms: (1) atom \(v\) stands for sampling a value \(v\); (2) trace \(\sigma\) stands for a nested trace \(\sigma\); (3) inj \(\sigma\) stands for branching, where \(\sigma\) traces the chosen branch; and (4) fold \(\sigma\) stands for a function call, where \(\sigma\) traces the function body.

For example, consider trace \(\sigma_{\text{main}}\) in (4.1) and its tree structure:

\[
\begin{align*}
\sigma_{\text{main}} & \triangleq \{ a : \text{fold } \sigma_0 \} \\
\sigma_0 & \triangleq \{ a : \text{atom } .7, b : \text{inj } \{d_1 : \text{fold } \sigma_1, d_2 : \text{fold } \sigma_2\} \} \\
\sigma_1 & \triangleq \{ a : \text{atom } .3, b : \text{inj } \{c : \text{atom } 2\} \} \\
\sigma_2 & \triangleq \{ a : \text{atom } .1, b : \text{inj } \{c : \text{atom } 9\} \}
\end{align*}
\]
Typing rules for the monadic fragment of FCC: terms, commands, and functions.

It is a trace of the PCFG model (Figure 5a). It is also a trace of the PCFG guide (Figure 5b), although the guide reorderings d1 and d2. Evaluating either program on this trace generates Add(Const(2), Const(9)) in the context-free language.

Contrast (4.1) with (2.1). Unlike the trace in (2.1), FCC traces do not stipulate a total ordering on events. An added benefit is that unsafe manual name mangling of the kind found in Figure 2a is avoided. That Pyro program would lead to name collisions and crash the runtime if not for manual name mangling such as L + "_C". By contrast, in FCC, RVs in trace σmain are uniquely identified even if they may correspond to the same term variable in the program text.

Figure 6 defines how terms and commands evaluate. Rules have forms \( \Delta; \Gamma \vdash e : \tau \) and \( \Delta; \Gamma \vdash m : \tau \# \Sigma \), meaning that on trace \( \sigma \) (resp. event \( s \)), command \( m \) (resp. term \( t \)) evaluates to \( v \) and produces weight \( w \). A command denotes a distribution over traces; weight \( w \) is the (unnormalized) density of the trace. In E:Sample, event atom \( v \) specifies the value \( v \) sampled, and the weight is adjusted by the distribution’s density at \( v \). In E:Branch, event \( \sigma \) specifies the nested trace \( \sigma \) on which to evaluate the chosen branch. In E:Call, event \( \sigma \) specifies the nested trace \( \sigma \) on which to evaluate the function body. E:Bnd:Term evaluates \( a = t ; m \) by splitting the trace into an event \( s \) and a smaller trace \( \{ a' : S' \} \) under which \( t \) and \( m \) are respectively evaluated. E:Bnd:Expr need not split the trace, as expression variables are not captured in traces. E:Observe performs conditioning, multiplying the weight by the density of the observed data.

4.3 Static Semantics of FCC

Typing judgments of the pure fragment of FCC have form \( \Delta; \Gamma \vdash e : \tau \). The rules are standard and omitted herein. \( \Delta \) are the environments of term variables and expression variables, respectively:

\[
\Delta ::= \bullet \mid \Delta, a : \tau \quad \Gamma ::= \bullet \mid \Gamma, x : \tau
\]

A distribution has type \( \text{dist}(\tau) \); samples from the distribution have type \( \tau \). For example, a Gamma distribution \( \text{Gamma}(e_1; e_2) \) has type \( \text{dist}(\mathbb{R}_+) \); the support is positive reals.

**Typing commands and terms.** Trace types (resp. event types) describe control-flow behaviors and probabilistic effects of commands (resp. terms). As Figure 4 shows, trace types \( \Sigma \) and event types \( S \)
recursive syntax. A term can sample from a distribution
arithmetic and logic operators.

4.1 Syntax of FCC

The monadic, probabilistic fragment of FCC consists of

4

data types

events

The monadic, probabilistic fragment of FCC consists of

4


an atom


typedef Fpcfg = {
    a : atom [0,1],
    b : {c : atom R} Const(c) +a<.5
    { d1 : Fpcfg, d2 : Fpcfg } Add(d1,d2)
} b

typedef Fpcfg = {•} b

a : atom [0,1]

b : {•} Const(c) +a<.5

{•} Add(d1,d2)

c : atom R

d1 : •

d2 : •

Figure 8. Trace type of the pcfg functions in Figure 5.

have mutually recursive syntax. The trace type of a command has form \{a : S\}_e, where a : S is an unordered map from term variables to event types, and e is the command’s return. Event types have four forms, corresponding to the four forms of events (or terms): atom τ, trace Σ, Σ₁ +ₜ Σ₂, and F(e).

Figure 7 defines the typing rules for commands and terms, which have forms Δ; Γ ⊢ₜ m : τ # Σ and Δ; Γ ⊢ₜ t : τ # S. They make sure that sampling, branching, and invocation of (possibly) recursive functions are recorded in trace types and event types. Figure 7 also gives the typing rule for global function definitions (S:DEF). Metavariable F ranges over names of functions’ trace types. For example, function pcfg in Figure 5a has trace type Fpcfg, defined recursively in Figure 8. The main command in Figure 5a then has trace type \{ a : Fpcfg \}_{unit}.

Two compatibly typed programs need not agree on the orders computations take place. In particular, it is possible for a compatibly typed guide to reorder computations under the same parent in a tree-structured trace type. For instance, the guide pcfg function in Figure 5b reorders subtrees d1 and d2, but it can still be typed as Fpcfg. The guide also does extra computations such as running neural networks. These computations are pure; they are not part of the guide’s trace type.

Checkpoints are part of trace types. Two compatibly typed programs must agree on control-flow decisions. Fpcfg captures the branching condition a < .5. So for the guide pcfg function to also have type Fpcfg, it should use a < .5 as the branching condition. More generally, two compatibly typed programs must agree on checkpoint expressions, which include branching condition e in \text{ite}(e; m₁; m₂), command return e in \text{ret}(e), and term argument e₁ in \text{call}(f; e₁; e₂). For example, trace type Fpcfg captures not only the branching condition a < .5 but also command returns such as {...}_\text{Const(c)} and {...}b. Consider the following command:

\begin{align*}
a &= \text{call}(pcfg); \\
b &= \text{ite}(\text{height}(a) < 6; m₁; m₂);
\end{align*}

If a model pcfg and its guide were allowed to disagree on command returns, then even though two programs use the same branching condition height(a) < 6, they could end up making different branching choices because they would disagree on a. Hence, to use trace typing to justify AC, the type system must keep track of checkpoints as part of trace types.

A consequence is that checkpoints can use only term variables—expression variables are not bound in trace types. In S:BRANCH, S:RETURN, and S:CALL, checkpoint expressions are type-checked under an empty environment • of expression variables. The requirement does not restrict the class of model programs that can be expressed, because the programmer can always use term variables and monadic returns to fix a program that would otherwise not type-check with expression variables. For example, the program below does not type-check because z, an expression variable, is used in the checkpoint z > 0:

\begin{align*}
a &= \text{sample}(\text{NORMAL}(x; y)); \\
z &= a + 1; \\
c &= \text{ite}(z > 0; m₁; m₂);
\end{align*}

This program is easily fixed by replacing z with a term variable b and replacing a + 1 with \text{ret}(a + 1).
It may appear restrictive that a guide must have syntactically identical checkpoints as its model. We found that in practice it is often trivial to modify manually created guides to meet this requirement. More importantly, it fits in a compiler pipeline where guides are automatically generated. Another benefit is that it allows typing to be entirely syntax-directed, thus striking a good balance between simplicity and expressivity.

There exist distributions whose supports depend on parameters. Like prior systems, the expressive power of FCC is limited by the lack of such dependently supported distributions in general. However, we believe that checkpoints offer a viable approach to incorporating them in the future. In particular, sampling from a delta distribution whose support is concentrated on $e^\theta$ can be encoded as $a = \text{ret}(e)$: because this checkpoint is part of the command's trace type, two programs are required to agree on $e$ to be compatibly typed.

**Term variables are more than binders.** Term variables in FCC are unusual, in the sense that they are not just binders, which have local scopes, but also labels, which can occur in trace types. Thus, renaming the term variables in a command will cause it to have a different trace type. In a trace type $\{a_1 : S_1, \ldots, a_N : S_N\}_e$, a term variable $a_i$ can occur in $e$ and the event types of its siblings.

We note that it is possible to adapt the FCC design so that binders and labels are decoupled, using a syntax like $x = \text{sample}(\ell; \text{Normal}(0; 1))$, where $x$ is a binder and $\ell$ a label. We choose to represent the binder and the label as one single term variable, for less verbiage.

### 4.4 Measure Semantics of FCC

**Definition 4.1.** A closed command $m$ of trace type $\Sigma (\bullet ; \bullet \vdash_m m : \tau \neq \Sigma)$ induces a measure $[m]$ over a measurable space $[\Sigma]$ of traces:

$$P_m(\sigma) \overset{\text{def}}{=} \begin{cases} w, & \text{if } \sigma \vdash_m m \Downarrow^w u, \\ 0, & \text{otherwise} \end{cases}$$

$[m] (A) \overset{\text{def}}{=} \int_A P_m(\sigma) \mu_\Sigma (d\sigma)$

Measure $[m]$ is defined by integrating a density function $P_m(\cdot)$ over a measurable set $A$ of traces. As is usual, only terminating executions have a positive density. Measure $[m]$ is in general un-normalized as $m$ may perform conditioning. When the normalization constant is in $(0, \infty)$, the normalized measure exists and is given by $\int P_m(\sigma) d\sigma / \int P_m(\sigma) d\sigma$.

The construction of the measurable space $[\Sigma]$ of traces, as well as that of its stock measure $\mu_\Sigma$, is by a double induction, first on a “step index” for handling recursive trace types, and then on the structure of trace types and event types. The definition can be found in Appendix A.2. We omit showing the measurability of $P_m(\cdot)$; we expect a proof to follow the technique found in Borgström et al. [2016] and Szymczak and Katoen [2019].

With the measure semantics for commands defined, we can formalize the notion of absolute continuity, which says the measure induced by one command is “supported” by that of the other:

**Definition 4.2.** A command $m_1$ is absolutely continuous with respect to another $m_2$, if $[m_1] (A) \neq 0$ on every measurable set $A$ of traces for which $[m_2] (A) \neq 0$.

### 4.5 Theoretical Results

Before we can establish the theorems, we introduce a notion of typing for traces and events, namely $s \vdash, S \Downarrow u$ and $\sigma \vdash, \Sigma \Downarrow^\sigma v$ defined in Figure 9. The typing rules evaluate checkpoint expressions in types to make sure that they agree with what happens in the trace. The evaluation happens in the pure fragment of FCC. Using the rules, we can derive that trace $\sigma_{\text{main}}$ from (4.1) is well-typed: $\sigma_{\text{main}} \vdash \{ a : \text{pcfg} \} \Downarrow \text{unit}$. 


We want a generated guide to faithfully yet parsimoniously respect conditional dependences. The Normalization (resp. event), then the value is well-typed and so is the trace (resp. event):

\[ \text{Theorems 4.3 and 4.4: a command } m \text{ of trace type } \Sigma \text{ has positive density } P_m(\sigma) > 0 \text{ on a trace } \sigma \text{ if and only if } \sigma \text{ has type } \Sigma. \text{ The guide generator need not type-check traces.} \]

However, the control-flow expressiveness of a universal PPL poses challenges.

5 IDENTIFYING CORRELATIONS AND INDEPENDENCES

We want a generated guide to faithfully yet parsimoniously respect conditional dependences. However, the control-flow expressiveness of a universal PPL poses challenges.
Consider conditional dependences in the PCFG model in Figure 5a. The model uses stochastic branching and general recursion, so it is impossible to express it as a BN. But given any of its traces, we can unfold recursion for that trace, as Figure 10 does. Nodes stand for returns of recursive calls to \texttt{pcfg}, and arrows signify dataflow. Absent of control flow, the dataflow graph is essentially a BN. Per the notion of BN active trails, colliders in Figure 10 mean that leaf nodes are correlated with each other and with \texttt{obs}. But we had to unroll recursion to identify such correlations for a given trace. Yet it is infeasible to unroll recursion for every possible trace, as general recursion yields traces of infinitely many shapes.

Adding to the complexities is that control dependences may or may not transmit correlation.

- In Figure 5a, we know that \texttt{a} (line 6) and \texttt{pcfg}'s return (line 14) must be correlated. We also know that this correlation is due to control dependence: \texttt{a} determines which branch is taken and thus influences \texttt{pcfg}'s return value.
- In contrast, we know that \texttt{a}'s value and \texttt{c}'s value (line 8) must be independent given \texttt{a} < .5. Yet \texttt{c} is control-dependent on \texttt{a}: \texttt{a} determines if \texttt{c} is sampled in the first place.

How can a guide generator identify such correlations and independences that involve control flow?

Fidelio extracts program dependences into a graph representation (Section 5.1). It uses a notion of active trails for the graph representation as an indicator of correlation (Sections 5.2 and 6).

### 5.1 A Graph Representation of Dependences

Program dependence graphs (PDGs) [Ferrante et al. 1987], employed by many optimizing compilers, are a classic idea to represent dependences between program elements. PDGs are appealing because they represent not only data dependences (as BNs do) but also control dependences. The exact definition of PDGs are tailored to specific languages and problems being considered.

Figure 11a shows two PDGs, one for the main command and the other for the \texttt{pcfg} function in Figure 5a. They are generated by the construction in Figure 12. While the presentation of Figure 12 is dense, the construction is mostly standard. It defines three sets \texttt{Nodes}, \texttt{CDEs}, and \texttt{DDEs} by structural induction. The sets contain nodes, control-dependence edges (CDEs), and data-dependence edges (DDEs) of the PDG.

There are two types of nodes in a PDG: control nodes and data nodes. Control nodes are drawn as rounded rectangles \(
\begin{array}{c}
\text{ENTRY} \\
\end{array}
\), one per function (or main command), and \(
\begin{array}{c}
\text{ ITE} \\
\end{array}
\), one per branching term \texttt{ite}(\texttt{e}; \texttt{m}_1; \texttt{m}_2). Metavariable \( C := \text{ENTRY} | \text{ITE} \) ranges over control nodes. Control nodes are sources of CDEs, which are drawn as gray, dashed arrows.

CDEs can have labels \( \ell ::= \top | \bot | \varepsilon \). CDEs from a branching node \( \text{ITE} \) are labeled either \( \top \) or \( \bot \), signifying the branch taken. For example, in Figure 11a, all CDEs from \( \text{ITE}(a < .5) \) are labeled. CDEs from \( \text{ENTRY} \) are not labeled (i.e., they have the null label \( \varepsilon \)). In Figure 12, the construction \( \text{CDE}_m(m \ell C) \) (resp. \( \text{DDE}_m(m \ell N) \)) defines CDEs induced by a command (resp. a term), where \( C \) is the current control parent, and \( \ell \) labels the CDE from the control parent.

Data nodes are drawn as rectangles \( \square \). They are created for variable bindings \( \texttt{a} \) and \( \texttt{obs} \), monadic returns \( \texttt{ret} \), call arguments \( \texttt{e} \), and conditionings \( \texttt{observe}(e_1; e_2) \). Data nodes are sources of DDEs, which are drawn as black, solid arrows.

In Figure 12, \texttt{DDE}_m(m \ell N) \texttt{(resp. DDE}_m(m \ell N) \texttt{)} defines DDEs induced by a command (resp. a term). To focus on terms in the monadic fragment, \texttt{Nodes}_m(m \ell N) does not create data nodes for expression variables. However, expression variables can transmit data dependence too. Thus, \texttt{DDE}_m(m \ell N) takes as input a substitution \( \gamma ::= \bullet | \gamma, x \mapsto e \) for expression variables occurring
For the PCFG model program, Fidelio uses the construction in Figure 12 to build the PDGs in (a) before generating the guide program. For comparison, guide PDGs are shown in (b). The PDG construction inserts the collider \( h \rightarrow \text{ret}(b) \), creating active trails between \( h \) and nodes in the PDG. Informed by the active trails, the guide generator makes sample and call terms depend on \( h \) (e.g., \( h \rightarrow a \) in (b)), thus faithfully expressing their correlation to historical information stored in \( h \) (Section 6).

**Figure 11.** For the PCFG model in Figure 5a and the PCFG guide in Figure 5b.

5.2 A Notion of Active Trails for PDGs

As Section 2 reviews, in BNs, absence of active trails implies conditional independence. We adapt the notion of active trails to PDGs. We do not claim soundness or completeness of this adapted notion in the general case, though it is sound and almost complete for BN-like programs. Soundness and completeness are theoretically appealing properties, but they require much more measure-theoretic sophistication and we leave them to future work. We satisfy ourselves with experimental evidence that absence (resp. presence) of active trails is a good indicator of independence (resp. correlation) (Figure 13), and with an extensive evaluation of how dependence awareness impacts the effectiveness of deep amortized inference (Section 8.2).

A definition of PDG active trails must take into consideration control dependences. Our definition uses a few terminologies. A control path, starting from \( \text{ENTRY} \), is a directed path consisting solely of CDEs. A control node \( C \) is a control ancestor of a PDG node \( N \) if \( C \) is on the control path to \( N \). Two nodes control-contradict each other if their control paths travel through mutually exclusive branches of the same branching node.

**Definition 5.1 (PDG active trails).** Let \( G \) be a PDG. Let \( M \) be a set of nodes in \( G \). Let \( N_1, N_2, \ldots, N_n \) be a trail in \( G \) (i.e., an undirected, acyclic path composed of CDEs and DDEs). The trail is said to be active given \( M \) if all the following conditions hold:

1. Control ancestors of \( N_1 \) and those of \( N_n \) are in \( M \).
2. \( M \) is closed under control ancestry. That is, control ancestors of any node \( M \in M \) are in \( M \), too.
Nodes (def f(a) = m) \equiv \{(ENTRY, [1]) \cup \text{Nodes}_{m}(m)\} # [1] and \square are synthetic nodes
Nodes_{m}(\text{ret}(e)) \equiv \{\text{ret}(e)\}
Nodes_{m}(\text{observe}(1_{1};1_{2}); m) \equiv \{\text{observe}(1_{1};1_{2})\} \cup \text{Nodes}_{m}(m)
Nodes_{m}(x = e; m) \equiv \text{Nodes}_{m}(m)
Nodes_{m}(a = t; m) \equiv \{a\} \cup \text{Nodes}_{m}(t) \cup \text{Nodes}_{m}(m)
Nodes_{m}(m) \equiv \text{Nodes}_{m}(m)
Nodes_{m}(\text{ite}(e; m_{1}; m_{2})) \equiv \{e\} \cup \text{Nodes}_{m}(m_{1}) \cup \text{Nodes}_{m}(m_{2})
Nodes_{m}(\text{sample}(e)) \equiv \emptyset
Nodes_{m}(\text{call}(f; e)) \equiv \emptyset

CDE_{m}(\text{def f(a) = m}) \equiv \text{CDE}_{m}(m) \in \{\text{ENTRY}\} # m has control parent \text{ENTRY}
CDE_{m}(\text{ret}(e)) \in C \equiv \{C \leftarrow \text{ret}(e)\}
CDE_{m}(\text{observe}(1_{1};1_{2}); m) \in C \equiv \{C \leftarrow \text{observe}(1_{1};1_{2})\} \cup \text{CDE}_{m}(m) \in C
CDE_{m}(x = e; m) \in C \equiv \text{CDE}_{m}(m) \in C
CDE_{m}(a = t; m) \in C \equiv \{C \leftarrow a\} \cup \text{CDE}_{t}(t) \in C \cup \text{CDE}_{m}(m) \in C
CDE_{m}(m) \in C \equiv \text{CDE}_{m}(m) \in C
CDE_{n}(\text{ite}(e; m_{1}; m_{2})) \in C \equiv \{C \leftarrow e\} \cup \text{CDE}_{m}(m_{1}) \cap \{e\} \cup \text{CDE}_{m}(m_{2}) \cap \{e\} # m_{1} has control parent \{e\}
CDE_{n}(\text{sample}(e)) \in C \equiv \emptyset
CDE_{n}(\text{call}(f; e)) \in C \equiv \emptyset

DDE_{m}(\text{def f(a) = m}) \equiv \{[1] \rightarrow \square\} \cup \text{DDE}_{m}(m) \bullet \square # \text{function's return flows into} \square
DDE_{m}(\text{ret}(e)) \in N \equiv \{[a] \rightarrow \text{ret}(e)\} | a \in \text{TV}_{Y}(e) \} \cup \{[\text{ret}(e)] \rightarrow \text{N}\}
DDE_{m}(\text{observe}(1_{1};1_{2}); m) \in N \equiv \{[a] \rightarrow \text{observe}(1_{1};1_{2})\} | a \in \text{TV}_{Y}(e_{1}) \cup \text{TV}_{Y}(e_{2}) \} \cup \text{DDE}_{m}(m) \in N
DDE_{m}(x = e; m) \in N \equiv \text{DDE}_{m}(m) (y, x \mapsto e) N \quad \# \text{substitution} y \text{ is extended}
DDE_{m}(a = t; m) \in N \equiv \text{DDE}_{t}(t) \in N \cup \text{DDE}_{m}(m) \in N \quad \# \text{evaluation result of} t \text{ flows into} \square
DDE_{m}(m) \in N \equiv \text{DDE}_{m}(m) \in N
DDE_{m}(\text{ite}(e; m_{1}; m_{2})) \in N \equiv \{[a] \rightarrow \text{ITE}(e; m_{1}; m_{2})\} | a \in \text{TV}_{Y}(e) \} \cup \text{DDE}_{m}(m_{1}) \in N \cup \text{DDE}_{m}(m_{2}) \in N
DDE_{n}(\text{sample}(e)) \in N \equiv \{a \mapsto \text{N} | a \in \text{TV}_{Y}(e)\}
DDE_{n}(\text{call}(f; e)) \in N \equiv \{a \mapsto \text{N} | a \in \text{TV}_{Y}(e)\} \cup \{e \mapsto \text{N}\}
\text{TV}_{Y}(e) \equiv \{a | a \in \text{FV}(e)\} \cup \{a | x \in \text{FV}(e) \land a \in \text{TV}_{Y}(x)\}

\text{Figure 12. Constructing PDGs for model programs.}

(3) No node in the trail control-contradicts \(N_{1}, N_{n}\), or \(M\).
(4) The trail is active given \(M\), in the sense of Bayesian networks. That is,

i) for any collider \(N_{i-1} \rightarrow N_{i} \leftarrow N_{i+1}\) in the trail, \(N_{i}\) or one of its descendants is either in \(M\) or a conditioning node (namely \(\text{observe(...)}\) or \(\square\));

ii) no other node in the trail is in \(M\) or is a conditioning node.

Only when two nodes are both reached in the same execution does it make sense to speak of correlation between them. Condition (1) ensures that when we speak of conditional dependence between \(N_{i}\) and \(N_{n}\), we do condition on the branching choices that caused \(N_{i}\) and \(N_{n}\) to be reached in the first place. Condition (2) further ensures that the branching nodes themselves are reached. Condition (3) says that correlation cannot be transmitted through nodes impossible to be reached when \(N_{1}, N_{n}\), and \(M\) are all reached. Finally, condition (4) requires that a PDG active trail be active in the BN sense.
Although the trail contains an entry and is considered. is inactive per condition (3): . is always in −→ c1, the trail −→ T in the (and CDEs from it) in PDG drawings hereafter. Also −→ L a is on the true branch. F a ←− F −→ ret ←− a observe −→ In is inactive in the BN sense, and so is the −→ −→ c1 entry a>0 observ ret observ d = sample(NORMAL(a;1)) ret(c1) else c2 = sample(NORMAL(a;1)) ret(c2) d = sample(NORMAL(0;4)) observe(NORMAL(b+d;0.1); obs) ret(unit) a = sample(NORMAL(0;2)) ret(c1) else c2 = sample(NORMAL(8;1)) ret(c2) d = sample(NORMAL(0;4)) observe(NORMAL(b+d;0.1); obs) ret(unit)

(a)

(b)

(c)

(d) Visualization of correlation and independence

Figure 13. Independence (resp. correlation) as visualized in (d) agrees with the absence (resp. presence) of PDG active trails per Definition 5.1. (ENTRY nodes have been elided in PDGs.)

By conditions (1) and (2), (ENTRY) is always in M, which by condition (4) implies that (ENTRY) cannot be on an active trail. So we elide (ENTRY) (and CDEs from it) in PDG drawings hereafter. Also notice that conditions (1)–(3) are automatically satisfied when the PDG does not have branching nodes—that is, when the PDG can be reduced to a BN. In this case, results about soundness [Verma and Pearl 1988] and (almost sure) completeness of active trails [Meek 1995] carry over.

We use examples in Figure 13 to further explain and justify Definition 5.1:

• An active trail can contain CDEs. In program (a), a and b are correlated, as is visualized by the top-left plot in (d). The correlation agrees with the presence of active trails between a and b in the PDG. All the active trails contain a CDE (e.g., [a −→ (a < 4) −→ [ret(c2) −→ b]]).

• Control ancestors of end nodes are always considered given. That is, they are in M per condition (1). Independence between a and c1 in (a) is visualized by the top-right plot in (d). In the PDG, because c1’s dependences are in question, the value of its control parent (a < 4) is considered known. Thus, given (a < 4), the trail [a −→ [a < 4] −→ c1] is inactive in the BN sense, and so is the trail [a −→ [a < 4] −→ c2 −→ [ret(c2) −→ b] −→ [ret(c1) −→ c1]].

• An active trail cannot contain nodes that control-contradict any end node. Program (b) differs from (a) in the line c2 = sample(NORMAL(a;1)). Hence, its PDG has an extra DDE [a −→ c2]. But still, a and c1 are independent, as visualized by the bottom-left plot in (d). The trail [a −→ c2 −→ [ret(c2) −→ b] −→ [ret(c1) −→ c1]] is inactive per condition (3): c2 and ret(c2) in the trail are on the false branch, whereas end node c1 is on the true branch.

• An active trail can contain intermediary nodes that control-contradict each other. In program (c), b and c are correlated, as visualized by the bottom-right plot in (d). The correlation agrees with an active trail in the PDG: [b −→ observe(...) −→ a < 0 −→ [observe(...) −→ c]]. Although the trail contains nodes on mutually exclusive branches, it is still active per Definition 5.1.
Active trails are compositional across call boundaries. Definition 5.1 does not unroll recursive functions. A call can create correlations internal to the callee; they are addressed by the callee’s PDG. A call can also create correlations across call boundaries. We can show that when a call is unrolled, a trail across the call boundary is active if and only if (1) the caller part of the trail is active and (2) the callee part of the trail is active. Below we illustrate this compositionalty using PDGs in Figure 11a:

- obs in the main command is correlated with c in the callee pcfg. The correlation is indicated by two active trails: \( \text{obs} \rightarrow\text{observe(...)\leftarrow a} \) in the caller, and \( \text{ret(b)}\rightarrow h\leftarrow \text{ret(\text{Const}(c))\leftarrow \text{c}} \) in the callee, with the call’s return value flowing from \( \text{ret(b)} \) to \( a \).

- Right subtree \( d2 \) is correlated with the \( c \) sampled by the left subtree \( d1 = \text{call(pcfg)} \), as indicated by two active trails: \( \text{d2} \rightarrow \text{ret(Add(d1,d2))\leftarrow d1} \) in the caller, and \( \text{ret(b)}\leftarrow h\leftarrow \text{ret(\text{Const}(c))\leftarrow c} \) in the callee, with the recursive call’s return value flowing from \( \text{ret(b)} \) to \( d1 \).

6 GUIDE GENERATION AS TYPE-DIRECTED, DEPENDENCE-AWARE TRANSLATION

We formalize guide generation as a translation of model programs. The translation makes sure that generated guides have the same trace type as their models, thereby guaranteeing absolute continuity. The translation can reorder computations under the same control parent and reconfigure data dependences using the notion of active trails, thereby offering faithfulness and parsimony.

Figure 14 presents the translation rules. The translation is by structural induction on typing derivations of commands, terms, and function definitions.

Translating function definitions. Compositionality of active trails (Section 5.2) implies that guide generation is intra-procedural and thus modular. A key insight is that guide functions can use an extra hidden-state parameter \( h \) as a proxy for transmitting correlations across call boundaries. Rule \( \text{GG:Def} \) defines the translation of a function, informed by its PDG \( G \). The translated function has an extra parameter \( h \).

At a call site, the guide generator gathers into a hidden state the historical information correlated with the call’s return, and passes the hidden state to the callee (\( \text{GG:Call} \)). In turn, inside the callee,\(^2\) if a sample or call term has an active trail to parameter \( h \) in the PDG—indicating correlation with history through the return value—then the guide generator makes that term depend on \( h \) (\( \text{GG:Sample} \) and \( \text{GG:Call} \)). We expand on this translation below.

Translating terms. Translation of a term \( t \) has form \( G; C; \Lambda'; h : R; a \vdash \langle \Delta; \Gamma; t : \tau \# S \rangle \rightsquigarrow t' \), where \( G \) is the PDG of the current function \( F \), \( C \) is the set of control ancestors, \( \Lambda' \) is the term variables in scope after reordering (done by \( \text{GG:Cmd} \)), \( h \) is the hidden-state parameter added for \( F \), and \( a \) is the term variable to which \( t \) and \( t' \) are bound. \( \text{GG:Branch} \) and \( \text{GG:Cmd} \) translate branching terms and nested commands, by recursively invoking command translation. \( \text{GG:Sample} \) and \( \text{GG:Call} \) translate sample and call terms, and they must rewire data dependences.

Translating sample terms by reconfiguring dependences. \( \text{GG:Sample} \) translates sample(\( e \)). It first chooses a learnable distribution \( D_{\phi_e} \) having the same type as \( e \). For instance, \( a = \text{sample(Unif)} \) in line 6 of Figure 5a is translated using \( \text{Beta}(\text{NN}_a(\cdot)) \) as the learnable distribution over \( R_{[0,1]} \) (lines 5–6 of Figure 5b). It is also possible to choose a universal density estimator using recent advances in normalizing flows [Huang et al. 2018].

\( \text{GG:Sample} \) then passes relevant data dependences \( \alpha_1, \ldots, \alpha_k \) as input to the learnable distribution via \( D_{\phi_e}(\alpha_1, \ldots, \alpha_k) \). Data dependences \( \alpha_i \in \{b_1, \ldots, b_n, h\} \) are chosen from variables lexically

\(^2\)Recall that in Figure 12, the PDG construction adds a synthetic node \( \text{d3} \) and connects it to the function’s return via a collider \( \text{d1} \rightarrow \text{d3} \leftarrow \text{ret(...)\leftarrow h} \). See, for instance, collider \( \text{d1} \rightarrow \text{d3} \leftarrow \text{ret(b)} \) in Figure 11a.
Given \( \epsilon \), for instance, consider translating \( \alpha \) to the learnable distribution \( \epsilon \rightarrow m' \)

\[
\mathcal{G}; \; C; \; \Delta'; \; h : \mathbb{R} 
\vdash \langle \Delta; \; \Gamma \vdash_m \; m : \tau \# \Sigma \rangle \quad \Rightarrow \quad m'
\]

(GG:CMD)

\[
\forall i. \; \mathcal{G}; \; C; \; \Delta', \; \alpha_i : \tau_{i_1}, ..., \alpha_{i_{N-1}} : \tau_{i_{N-1}}; \; h : \mathbb{R}; \; \alpha_i \vdash \langle \Delta_i; \; \Gamma_i \vdash_{t_i} \; \tau_{i_1} \# S_{i_1} \rangle \quad \Rightarrow \quad t_i'
\]

\[
\mathcal{G}; \; C; \; \Delta'; \; h : \mathbb{R}; \; a \vdash \langle \Delta; \; \Gamma \vdash_t \; t : \tau \# S \rangle \quad \Rightarrow \quad t'
\]

(GG:TM:CALL)

\[
M \overset{\text{def}}{=} C \cup \{ \langle \mathbb{R} \rangle, \langle \mathbb{H} \rangle \} \quad \text{there is an active trail in } \mathcal{G} \text{ between } [a] \text{ and } [ \mathbb{H}]; \; \text{given } \mathcal{M} - \{ [ \mathbb{R} ] \}
\]

\[
\mathcal{G}; \; C; \; b_1 : \tau_1, ..., b_n : \tau_n; \; h : \mathbb{R}; \; a \vdash \langle \Delta; \; \Gamma \vdash \text{call}(f; e) : \tau \# F(e) \rangle \quad \Rightarrow \quad \text{call}(f; e; \text{op}_\phi(\alpha_1, ..., \alpha_k))
\]

(GG:TM:BRANCH)

\[
\forall i. \; \mathcal{G}; \; C \cup \{ [ \mathbb{C} ] \} \quad \Delta'; \; h : \mathbb{R} \vdash \langle \Delta; \; \Gamma \vdash_{m_1} \; m_1 : \tau \# \Sigma_1 \rangle \quad \Rightarrow \quad m'_i
\]

\[
\mathcal{G}; \; C; \; \Delta'; \; h : \mathbb{R}; \; a \vdash \langle \Delta; \; \Gamma \vdash \text{ite}(e; m_1; m_2) : \tau \# \Sigma_1 + r \Sigma_2 \rangle \quad \Rightarrow \quad \text{ite}(e; m'_1; m'_2)
\]

(GG:TM:CMD)

\[
\mathcal{G}; \; C; \; \Delta'; \; h : \mathbb{R}; \; a \vdash \langle \Delta; \; \Gamma \vdash \text{call}(f; e) : \tau \# F(e) \rangle \quad \Rightarrow \quad \text{call}(f; e; \text{op}_\phi(\alpha_1, ..., \alpha_k))
\]

(GG:Def)

\[
f \vdash \langle \tau_1 \rangle \quad \Rightarrow \quad \tau_2 \# F \quad \text{typedef } F = \forall a : \tau_1. \; \Sigma
\]

\[
\mathcal{G}; \; \{ [\text{ENTRY}] \}; \; a : \tau_1; \; h : \mathbb{R} \vdash \langle [a : \tau_1] \bullet \vdash_m \; m : \tau_2 \# \Sigma \rangle \quad \Rightarrow \quad m'
\]

\[
\mathcal{G}; \; h : \mathbb{R} \vdash \langle \text{def } f(a) = m : \langle \tau_1 \rangle \quad \Rightarrow \quad \tau_2 \# F \rangle \quad \Rightarrow \quad \text{def } f(a)(h) = m'
\]

Figure 14. Guide generation as translation of typing derivations

in scope, including the hidden-state parameter \( h \). The translation uses existence of PDG active trails as an indicator of correlation. That is, \( \alpha_i \) is considered a dependency of \( a = \text{sample}(D_\phi(\ldots)) \) if an active trail exists in \( \mathcal{G} \) between \( [ \alpha_i ] \) and \( [a] \) given all other variables in scope (as well as all control ancestors).

For instance, consider translating \( c = \text{sample}(\text{NORMAL}(0; 1)) \) in line 8 of Figure 5a. In the PDG (Figure 11a), trail \( [c \rightarrow \text{ret}(\text{Const}(0)) \rightarrow [h] \rightarrow \text{ret}(b) \rightarrow \rightarrow \rightarrow h] \) is active given \( \{ [\text{ENTRY}], [a < \mathbb{S}]. \} \}. \) So \( h \) is input to the learnable distribution \( \text{NORMAL}(\text{NN}_{\mathbb{C}}(\cdot)) \) for \( c \) in lines 8–9 of Figure 5b:

\[
\mathcal{G}; \; \{ [\text{ENTRY}], [a < \mathbb{S}] \}; \; a : \mathbb{R}_{[0,1]}; \; h : \mathbb{R}; \; c \vdash \langle a : \mathbb{R}_{[0,1]}; \; \bullet \vdash \text{sample}(\text{NORMAL}(0; 1)) : \mathbb{R} \# \text{atom } \mathbb{R} \rangle
\]

\[
\Rightarrow \text{sample}(\text{NORMAL}(\text{NN}_{\mathbb{C}}(h)))
\]
Translating call terms by evolving hidden states. GG:TM:CALL translates function calls. As in GG:TM:SAMPLE, dependencies \( \alpha_1, ..., \alpha_k \) are those in-scope variables to which active trails exist in the PDG. They are input to a learnable function \( \Omega_{\phi_a}(\cdot) \), whose output is then passed to the callee as its hidden-state parameter.

For instance, consider translating \( d_1 = \text{call}(\text{pcfg}) \) in line 11 of Figure 5a. In the PDG,

- \( \text{ret}(\text{Add}(d_1, d_2)) \rightarrow d_2 \) is an active trail given \( \{ \text{ENTRY}, (a < 5), [1], [a] \} \), and

- \( d_1 \rightarrow \text{ret}(\text{Const}(c)) \rightarrow h \rightarrow \text{ret}(b) \rightarrow b \rightarrow h \) is an active trail given \( \{ \text{ENTRY}, (a < 5), d_2, [a] \} \).

So both \( h \) and \( d_2 \) are merged into the hidden state passed to the callee, in lines 14–15 of Figure 5b:

\[
G: \{ \text{ENTRY}, (a < 5) \}; a : \mathbb{R}_{[0,1]}, d_2 : \text{Expr}; h : \mathbb{R}; d_1 \vdash_1 (a : \mathbb{R}_{[0,1]}; \bullet \vdash_1 \text{call}(\text{pcfg}) : \text{Expr} \# F_{\text{pcfg}}) \\
\approx \text{call}(\text{pcfg}; \text{NN}_{d_1}(h, \text{embed}(d_2)))
\]

\( d_2 \) needs an embedding layer because it is a syntax tree (of type Expr) rather than a scalar value.

If we unfold recursion in the PCFG guide for the same trace of execution as in Figure 10, we get Figure 15a, which makes it explicit how the hidden state is evolved to include history. Initially, the hidden state contains only information about \( \text{obs} \). Visiting a right subtree simply uses the parent’s hidden state. Before visiting a left subtree \( d_1 \), the hidden state is updated to include information about the right subtree \( d_2 \) just visited (see the leftward horizontal arrows \( h \rightarrow b \)).

While information about historically sampled \( c \)’s flows into the hidden state via returns of recursive calls \( b \), information about historically sampled \( a \)’s is never merged into \( h \). As the beginning of Section 5 analyzes, an \( a \) is independent of all RVs to be sampled later in the current subtree, given the branching condition \( a < 5 \). An \( a \) is also independent of all RVs to be sampled after the current call returns, given the call’s return value \( b \).

Contrast that with Figure 15b, which visualizes dataflow for the same trace but in a guide that uses an RNN to create correlations [Le et al. 2017]. Unaware of the program dependence structure, the RNN inevitably renders all RVs correlated. As the syntax tree is traversed, the RNN’s hidden state accretes information about every RV visited thus far, causing a later RV to treat all earlier RVs as dependencies. As a result of the to-and-fro dataflow, hidden states contain unneeded information, diluting learning signals.

Translating commands and reordering terms. Rule GG:CMD translates a model command, which has the trace type \( \{ a_1 : S_1, ..., a_N : S_N \}_e \), to a guide command of form \( a_i = t_i; ... ; a_N = t_N; \text{ret}(e) \).

Importantly, the guide command can choose a different ordering \( \{ a_j = t_j \}_{j=1}^N \) of term bindings than the original ordering \( \{ a_j = t_j \}_{j=1}^N \). Rule REORDER specifies the constraint that reordering must satisfy. In particular, if \( a_{i_j} \) is a free variable in a checkpoint node in term \( t_k \), then \( a_{i_j} = a_{i_j} \) must occur before \( a_{i_k} = t_k \). Recall from Section 4.3 that checkpoints are part of trace types, so satisfaction of
the constraint helps ensure that the translated guide command can use the same checkpoints and thus be compatibly typed.

Each subterm of the model command is translated by recursively invoking the term translation rules. The term-variable environment for subterm $t'_i$ is extended with $a_{I_1}, \ldots, a_{I_i-1}$, the term variables in scope for the guide term after reordering.

Reordering of term bindings is often desirable because it can lead to fewer dependences while maintaining faithfulness. Fidelio follows the heuristic of Stuhlmüller et al. [2013] and Paige and Wood [2016] to invert topological orderings (while respecting the reordering constraint).

Consider a model that has a PDG as shown in Figure 16a. There, $a_1, a_2, a_3, a_4, b_1, b_2, c$ is a valid topological ordering of the term bindings. In (b) is the PDG of a faithful guide, using the same ordering. (Notice that removing any DDE makes it unfaithful by introducing conditional independences not found in the model.) In (c) is the PDG of another faithful guide, using an inverted topological ordering $c, b_1, b_2, a_1, a_2, a_3, a_4$. Whereas (b) is almost fully connected, guide (c) attains a sparse dependence structure using available conditional independences—e.g., given $b_2$, $a_3$ is independent of all variables in \{obs, c, b_1, a_1, a_2\}. As we show in Section 8, for the same total number of neural-network parameters, guide (c) leads to better performance than guide (b), both for training and for inference.

**AC by construction.** We prove that the translation preserves trace typing:

**Theorem 6.1 (Type-Preserving Translation).**

If $G; \{\text{ENTRY}\}; \bullet; \text{obs} : \mathbb{R} \vdash m_m : \tau \# \Sigma \Rightarrow m_g, \text{then} \bullet; \text{obs} : \mathbb{R} \vdash m_g : \tau \# \Sigma$.

It follows from Theorems 4.5 and 6.1 that for a well-typed model, the translation is guaranteed to generate a guide that is mutually absolutely continuous with the model:

**Theorem 6.2 (Absolute Continuity by Construction).**

If $G; \{\text{ENTRY}\}; \bullet; \text{obs} : \mathbb{R} \vdash m_m : \tau \# \Sigma \Rightarrow m_g \text{ and } v_{\text{obs}} : \mathbb{R}$, then for any measurable set $A$ of traces, $\| m_m \{v_{\text{obs}}/\text{obs}\} \| (A) \neq 0 \text{ if and only if } \| m_g \{v_{\text{obs}}/\text{obs}\} \| (A) \neq 0$.

7 IMPLEMENTATION

Fidelio uses Pyro [Bingham et al. 2019] as the underlying inference engine. A Fidelio model is compiled to a model and a guide in Pyro with explicit name mangling. The user can specify as arguments hyperparameters including the number of layers in a network, the number of neurons in each layer, and the activation function used. It is also easy to modify neural networks in generated guides, as we do in setting up our experiments (Section 8.2).

A few surface-syntax features help streamline guide generation. (1) Fidelio supports directives using which the programmer can hint shapes of tensors. (2) Fidelio supports directives for specifying embedding functions for data that warrant special embedding layers (e.g., convolutional networks.
Table 1. ● model and guide have compatible types. ○ model and guide fail to be compatibly typed. Last seven programs appeared in the expressiveness evaluation conducted by Wang et al. [2021].

<table>
<thead>
<tr>
<th>Program</th>
<th>Description</th>
<th>Lew et al. [2019]</th>
<th>Wang et al. [2021]</th>
<th>Fidelio</th>
</tr>
</thead>
<tbody>
<tr>
<td>collider</td>
<td>Figure 1 (model &amp; guide2)</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>pcfg</td>
<td>Figure 5, (a) &amp; (b)</td>
<td>○</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>treebn</td>
<td>Figure 16, (a) &amp; (c)</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>curvefit</td>
<td>Curve fitting, Lew et al. [2019, Fig. 4]</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>kalman</td>
<td>Kalman Smoother</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>branching</td>
<td>Random Control Flow</td>
<td>○</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>marsaglia</td>
<td>Marsaglia Algorithm</td>
<td>○</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>aircraft</td>
<td>Aircraft Detection</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>gmm</td>
<td>Gaussian Mixture Model</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>sprinkler</td>
<td>Bayesian Network</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>dp</td>
<td>Dirichlet Process</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
</tbody>
</table>

as image embeddings). (3) Fidelio provides an option that generalizes MADE [Germain et al. 2015] to allow weight sharing among local networks.

8 EVALUATION

8.1 Expressiveness of the Trace-Type System

Table 1 evaluates the expressiveness of Fidelio’s type system, comparing it with state-of-the-art type systems for AC [Lew et al. 2019; Wang et al. 2021]. The benchmark programs include examples in this paper and also those taken from prior work [Lew et al. 2019; Wang et al. 2021], Anglican [Wood et al. 2014], and Pyro, with reasonable modifications so that they conform to our syntax.

The type system of Lew et al. [2019] cannot type-check models of pcfg, branching, and marsaglia, because they use branching and recursion in their general forms. The type system of Wang et al. [2021] cannot type-check guides of collider, pcfg, treebn, and curvefit, because they reorder computations. dp uses stochastic memoization, a PPL feature not yet supported by these systems.

8.2 Performance Implication of Dependence-Awareness for Training and Inference

We assess the implications of adopting dependence-aware guide generation for both ahead-of-time training and run-time inference. Table 2 summarizes our benchmarks, which include a variety of applications, program sizes, dependence structures, and training setups.

We compare Fidelio mainly with two methods for guide generation in a universal PPL: mean-field guides, which make strong independence assertions, and RNN guides, which make all variables correlated. Generated Fidelio guides were modified so that they have the same number of trainable parameters (i.e., capacity) as the other two guides. Networks in the Fidelio guides and mean-field guides typically have no more than four linear layers and use ReLU activation. All guides for a benchmark program are trained for the same number of iterations. An appendix in the technical report documents experimental setups and additional statistics. The RNN we use in all experiments is LSTM [Hochreiter and Schmidhuber 1997], which deals with the vanishing gradient problem and is thus better at remembering correlation than vanilla RNNs.

Table 2 shows validation loss as a metric for evaluating training performance. Fidelio consistently leads to lower validation loss as computed by (2.3)—trained Fidelio guides are closer approximations to true posteriors than mean-field or LSTM guides, averaged on the validation sample set.

Validation losses alone do not tell the whole story, however. Below we explain our benchmarks and findings in more detail.
Table 2. Benchmarks for evaluating different approaches to guide generation. Column "#RVs": expected number of RVs sampled in a trace. Validation loss is computed per (2.3) on a validation sample set.

<table>
<thead>
<tr>
<th>Program</th>
<th>Structures</th>
<th>#RVs</th>
<th>LoC (model)</th>
<th>Capacity</th>
<th>Validation Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Mean-Field</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LSTM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Fidelio</td>
</tr>
<tr>
<td>treebn</td>
<td>BN</td>
<td>31</td>
<td>37</td>
<td>2K</td>
<td>51.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>45.76</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>41.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>51.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>43.96</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>41.04</td>
</tr>
<tr>
<td>ecoli70</td>
<td>BN</td>
<td>46</td>
<td>54</td>
<td>10K</td>
<td>11.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>15.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>16.56</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>11.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>15.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>16.60</td>
</tr>
<tr>
<td>longrange (k = 5)</td>
<td>loop</td>
<td>8</td>
<td>16</td>
<td>1.4K</td>
<td>25.39</td>
</tr>
<tr>
<td>longrange (k = 10)</td>
<td>loop</td>
<td>13</td>
<td>16</td>
<td>1.4K</td>
<td>44.05</td>
</tr>
<tr>
<td>longrange (k = 20)</td>
<td>loop</td>
<td>23</td>
<td>16</td>
<td>1.4K</td>
<td>81.23</td>
</tr>
<tr>
<td>longrange (k = 40)</td>
<td>loop</td>
<td>43</td>
<td>16</td>
<td>1.4K</td>
<td>155.8</td>
</tr>
<tr>
<td>captcha</td>
<td>recursion</td>
<td>6</td>
<td>43</td>
<td>4.4M</td>
<td>1.644</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>−1.325</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>−1.212</td>
</tr>
<tr>
<td>astropcfg</td>
<td>recursion</td>
<td>15.0</td>
<td>159</td>
<td>1.1M</td>
<td>7.962</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>14.84</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.603</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6.226</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.352</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.980</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6.217</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.362</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.035</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6.216</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.387</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.114</td>
</tr>
<tr>
<td>mathcaptcha</td>
<td>recursion</td>
<td>14.5</td>
<td>237</td>
<td>1.25M</td>
<td>8.150</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>14.78</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.242</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8.329</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>14.79</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.950</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.830</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9.531</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.511</td>
</tr>
<tr>
<td>gmmcc</td>
<td>recursion</td>
<td>217.4</td>
<td>49</td>
<td>20K</td>
<td>111.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>257.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>16.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>111.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>170.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>15.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>111.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>152.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>14.55</td>
</tr>
</tbody>
</table>

Long-range correlation. Benchmark longrange [Harvey et al. 2019] tests the ability for guides to capture correlation between variables sampled far apart. The model samples two Gaussian variables $a$ and $b$—and also $k$ noise variables in between—before observing the sum of all $k + 2$ variables. We assess how well the inferred posteriors capture the strong correlation between $a$ and $b$ sampled $k$ variables apart, with $k \in \{5, 10, 20, 40\}$. We draw 2,000 samples after training each guide for 20,000 steps and plot them in Figure 17. In columns one and three, samples are drawn from the trained guides. In columns two and four, samples are drawn using IS, with the guides as proposals. The LSTM guide manages to capture the correlation to some degree when $k = 10$. However, the LSTM guide fails to remember correlation when $k = 20$. Quality of the learned proposal distributions has a knock-on effect on IS. When $k = 20$, the LSTM guide results in visibly lower-quality IS samples than the Fidelio guide of equal capacity. We also compute the effective sample size (ESS) for each guide, as a quantitative (though not comprehensive) metric of sample quality [Kong 1992]; Fidelio guides lead to significantly higher effective sample sizes.

Computation reordering. Benchmark treebn assesses the impact of computation reordering on training and inference performance. The model is linear Gaussian BN, similar to Figure 16a in structure. In addition to the mean-field and LSTM guides, we compare two Fidelio guides that either retain or invert the model’s topological ordering of variables (cf. Figures 16b and 16c). Figure 18 plots the training loss (2.3) as training progresses. All four guides have the same capacity (5K). The sparser Fidelio guide converges to the best solution. The table in Figure 18 shows the Fidelio guide with an inverted topological ordering results in significantly higher ESS than all the other guides: a sparser yet faithful dependence structure makes inference more effective.
The sparser guide is rejected by a prior type system [Wang et al. 2021], which disallows reordering computations. Trace-type systems have performance implications.

**ECOL170: a large Bayesian network.** We use benchmark ecoli70 [Schäfer and Strimmer 2005] to evaluate the implication of dependence-aware guide generation for large BN models. The program is a linear Gaussian BN with 46 variables and 70 arcs. Results can be found in an appendix; they show (1) that the mean-field guide and the LSTM guide did equally poorly on ecoli70 and (2) that Fidelio significantly improves on these dependence-agnostic approaches.

**Captcha: a regular grammar.** In benchmark captcha [Mansinghka et al. 2013], the model generates a captcha image by sampling a probabilistic regular grammar, and the inference problem is to recognize captcha text given an image. Guide programs use convolutional networks as image embeddings (Section 7). Experimental results can be found in an appendix.

While the mean-field guide did poorly on captcha, the LSTM guide and the Fidelio guide did almost equally well, in terms of convergence and recognition rates. A possible explanation is given by the structure of the model. Because captcha samples from a regular—rather than context-free—language, the dataflow exhibits a linear-recursive pattern (shown in an appendix). As a result, the model has almost no conditional independence for Fidelio to exploit, and the unfolded dataflow graph for the Fidelio guide is almost identical to that induced by an RNN.

**Astronomer: a context-free grammar.** In benchmark astropcfg [Manning and Schütze 1999], the model samples sentences such as “astronomers saw stars with telescopes”. We consider the inference task of regenerating observed sentences. The program consists of mutually recursive functions corresponding to nonterminal symbols in the PCFG.

Figure 19 shows the training-loss profile for all guides of the same capacity (175K). It also shows inference accuracy, as measured by sentence regeneration rates and Levenshtein edit distances for
all guides and for three capacities. Fidelio demonstrates a clear advantage over the LSTM guide. In particular, it increases regeneration rate by 10% on average. The results suggest that it pays off to exploit the tree-recursive dataflow pattern and the resulting conditional independences in the astropcfg model (Figures 10 and 15).

Math captcha. We design a second captcha benchmark, mathcaptcha, that samples simple arithmetic expressions and renders them as captcha images (e.g., $2+2+6+1$). Unlike captcha, mathcaptcha samples from a context-free language, so we expect Fidelio to improve convergence and captcha recognition rate over the LSTM guide.

Figure 20 confirms our belief: Fidelio outperforms mean-field and LSTM guides, by a large margin. The LSTM guide did not start to pick up until network capacity was raised to 3M (including an invariant 1M for a CNN embedding). In contrast, with limited network capacity and training budget, Fidelio can do better. Conditional independences inherent in PCFG-like models make them favored applications for Fidelio.

Gaussian mixture model for clustering and classification. In benchmark gmmcc [Le et al. 2017], the model samples a stochastic number of Gaussians and further samples data points from
them. Given an observed dataset, the inference task is to simultaneously identify clusters and
classify data points into clusters. This model is challenging because it samples a large number
of RVs (~200). Nonetheless, since data points are independently distributed, it offers conditional
independences that Fidelio can potentially tap into.

Figure 21 plots the training-loss profile and kernel density estimations: with the same number
of particles, the Fidelio guide is the only guide that correctly inferred the number of clusters.

Summary. We observe that Fidelio consistently improves over LSTM and mean-field guides, both
on convergence of training and on accuracy of inference. Our experiments suggest that compared
with LSTM, it is most favorable to use Fidelio for models that contain conditional independences
and contain ~10 RVs or more.

For models containing ~20 RVs or more, a mean-field guide may be competitive against LSTM;
see ecoli70, longrange (k = 20, 40), mathcaptcha, and gmmcc. Unaware of program dependence
structures, LSTM has to accrete information about all RVs in its hidden state. Unlike Fidelio or
mean-field guides, LSTM guides are unable to directly express any conditional independence, which
leads to diluted learning signals especially when the number of RVs is large. In contrast, Fidelio
guides capitalize on available conditional independences, so training can focus on learning the real
correlations.

9 RELATED WORK

Language-based solutions to absolute continuity. Lew et al. [2019] and Wang et al. [2021]
design type systems to check AC statically. Lew et al. [2019] formulate AC in terms of a denotational
measure semantics for a trace-based PPL. Their type system allows some forms of stochastic control
flow, including loops, but disallows general recursion or the kind of stochastic branching needed
for models such as PCFGs. It allows guides to reorder computations in some way.

Wang et al. [2021] propose a coroutine-based PPL, where a model and a guide send and receive
messages over communication channels to synchronize on sampled values, branching choices,
and function calls. AC is formulated in terms of a measure semantics induced by this operational
semantics. The approach is inspired by session types [Honda et al. 1999]: their guide types describe
communication protocols of model–guide pairs. Unlike Lew et al. [2019], the type system allows
branching and recursion in their general forms. However, it requires that computations happen in
exactly the same (total) order as prescribed by guide types.

In both prior approaches, it is the programmer who authors guides and makes sure that guides
have compatible types with models. In our approach, guides are automatically generated and are
guaranteed to have compatible types with models.

Lee et al. [2019] introduce a static analysis for Pyro’s stochastic variational inference. The
analysis aims to prove that a guide has the same support as the model and satisfies differentiability
conditions. It deals with Pyro features including tensors and plating, but it does not support
stochastic branching or general recursion.

Generating guides automatically. A recent spate of work in machine learning has focused on
generating guide programs [Stuhlmüller et al. 2013; Paige and Wood 2016; Le et al. 2017; Webb
et al. 2018, 2019; Baudart and Mandel 2021]. Except for Le et al. [2017], the prior approaches
deal with graphical models or non-universal PPLs; stochastic branching and general recursion
are not supported. Le et al. [2017] address the challenge by having a guide program talk to an
LSTM network, leading to faithful but “fully connected”, non-parsimonious guides. Webb et al.
[2018] create faithful, parsimonious guides, but only for BNs. Webb et al. [2019] and Weilbach et al.
[2020] integrate normalizing flows [Rezende and Mohamed 2015; Huang et al. 2018] to increase the
representational power of learnable continuous distributions; it is an orthogonal extension that can be incorporated into Fidelio.

**Dependence analysis for PPLs.** Hur et al. [2014] study program slicing for an imperative PPL with loops. Like guide generation, slicing takes into account conditional dependences as indicated by active trails. Unlike guide generation, the slicer is concerned with conditional dependences of a program’s return value, which exists in all finite executions, whereas a guide generator is concerned with conditional dependences of sample and call terms, which are reached only stochastically and subject to reordering in guides. Gorinova et al. [2021] use an information-flow type system to identify conditional independences in a non-universal PPL. It is interesting to study if the information-flow perspective can be applied to a universal PPL, which could serve as a good basis for studying soundness of PDG active trails as proposed in Section 5.2.

## 10 CONCLUSION

We have presented Fidelio, a framework that can automatically generate guide programs for deep amortized inference in a universal PPL. It addresses challenges posed by the control-flow expressiveness of a universal PPL. Fidelio uses a novel trace-type system and generates guides via a type-guided, dependence-aware translation of models. Theoretical results show that the translation is type-preserving and thus ensures absolute continuity by construction. Experimental results show that **symbolic** dependence information in probabilistic programs can be used effectively to aid in **neural**-network-based inference.

## ACKNOWLEDGMENTS

We thank the anonymous reviewers for their valuable feedback. We thank Nada Amin, Hongfei Fu, Edward Lee, Yingzao Li, Di Wang, and Rob Zinkov for discussions and help. This work was supported by the Natural Sciences and Engineering Research Council of Canada. The views and opinions expressed are those of the authors and do not necessarily reflect the position of any funding agency.

## REFERENCES


Synthesizing Guide Programs for Sound, Effective Deep Amortized Inference

arXiv:2010.11887


A COMPLETE SEMANTICS OF FCC

In this appendix, we present the complete semantics of FCC and proofs of the main theorems.

A.1 Operational and Static Semantics

Figure 23. Evaluation rules for terms and commands.
\[ \Delta; \Gamma \vdash e : \tau \]

<table>
<thead>
<tr>
<th>Rule</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(TE:EVar)</td>
<td>( \Delta; \Gamma, x : \tau \vdash x : \tau )</td>
</tr>
<tr>
<td>(TE:TVar)</td>
<td>( \Delta; \Gamma \vdash a : \tau )</td>
</tr>
<tr>
<td>(TE:Triv)</td>
<td>( \Delta; \Gamma \vdash \text{unit} : \text{One} )</td>
</tr>
<tr>
<td>(TE:False)</td>
<td>( \Delta; \Gamma \vdash \text{false} : \tau )</td>
</tr>
<tr>
<td>(TE:Cond)</td>
<td>( \Delta; \Gamma \vdash e : 2 )</td>
</tr>
<tr>
<td>(TE:FNat)</td>
<td>( \Delta; \Gamma \vdash \text{if}(e; e_1; e_2) : \tau )</td>
</tr>
<tr>
<td>(TE:Nat)</td>
<td>( \Delta; \Gamma \vdash n : \mathbb{N} )</td>
</tr>
<tr>
<td>(TE:App)</td>
<td>( \Delta; \Gamma \vdash e_1 : \tau_1 \rightarrow \tau_2 )</td>
</tr>
<tr>
<td>(TE:Bere)</td>
<td>( \Delta; \Gamma \vdash \text{Bern}(e) : \text{dist}(2) )</td>
</tr>
<tr>
<td>(TE:Let)</td>
<td>( \Delta; \Gamma \vdash \text{let}(e_1; x; e_2) : \tau_2 )</td>
</tr>
<tr>
<td>(TE:Gamma)</td>
<td>( \Delta; \Gamma \vdash e_1 : \mathbb{R}_+ )</td>
</tr>
<tr>
<td>(TE:Normal)</td>
<td>( \Delta; \Gamma \vdash \text{Normal}(e_1; e_2) : \text{dist}(\mathbb{R}) )</td>
</tr>
<tr>
<td>(TE:Geo)</td>
<td>( \Delta; \Gamma \vdash \text{Geo}(e) : \text{dist}(N) )</td>
</tr>
<tr>
<td>(TE:Geo)</td>
<td>( \Delta; \Gamma \vdash e : \mathbb{R}_+ )</td>
</tr>
<tr>
<td>(TE:Beta)</td>
<td>( \Delta; \Gamma \vdash \text{Beta}(e_1; e_2) : \text{dist}(\mathbb{R}[0,1]) )</td>
</tr>
<tr>
<td>(TE:Bere)</td>
<td>( \Delta; \Gamma \vdash \text{Bern}(e) : \text{dist}(2) )</td>
</tr>
<tr>
<td>(TE:Beta)</td>
<td>( \Delta; \Gamma \vdash \text{Beta}(e_1; e_2) : \text{dist}(\mathbb{R}[0,1]) )</td>
</tr>
<tr>
<td>(TE:Normal)</td>
<td>( \Delta; \Gamma \vdash \text{Normal}(e_1; e_2) : \text{dist}(\mathbb{R}) )</td>
</tr>
<tr>
<td>(TE:Geo)</td>
<td>( \Delta; \Gamma \vdash \text{Geo}(e) : \text{dist}(N) )</td>
</tr>
<tr>
<td>(TE:Geo)</td>
<td>( \Delta; \Gamma \vdash e : \mathbb{R}_+ )</td>
</tr>
</tbody>
</table>

Figure 24. Typing rules for expressions and values.
\[ \Delta; \Gamma \vdash_\Delta t : \tau \# S \]

(S:CMD)
\[ \Delta; \Gamma \vdash_\Delta m : \tau \# \Sigma \]
\[ \Delta; \Gamma \vdash_\Delta m : \tau \# \text{trace} \Sigma \]

(S:RETURN)
\[ \Delta; \bullet \vdash_\Delta e : \tau \]
\[ \Delta; \Gamma \vdash_\Delta \text{ret}(e) : \tau \# \{ e \} \]

(S:观察)
\[ \Delta; \Gamma \vdash_\Delta \text{observe}(e_1; e_2) : m : \tau \# \Sigma \]

(S:Branch)
\[ \Delta; \bullet \vdash_\Delta e : 2 \]
\[ \Delta; \Gamma \vdash_\Delta m_1 : \tau \# \Sigma_1 \]
\[ \Delta; \Gamma \vdash_\Delta m_2 : \tau \# \Sigma_2 \]

(S:Call)
\[ f : \langle \tau_1 \rangle (\tau_2) \leadsto \tau_3 \# F \]
\[ \Delta; \bullet \vdash_\Delta e_1 : \tau_1 \]
\[ \Delta; \Gamma \vdash_\Delta e_2 : \tau_2 \]

\[ \Delta; \Gamma \vdash_\Delta \text{call}(f; e_1; e_2) : \tau_3 \# F(e_1) \]

\[ \Delta; \Gamma \vdash_\Delta m : \tau \# \Sigma \]

(S:Sample)
\[ \Delta; \Gamma \vdash_\Delta e : \text{dist}(\tau) \]

(S:Bnd:Expr)
\[ \Delta; \Gamma \vdash_\Delta e : \tau_1 \]
\[ \Delta; \Gamma \vdash_\Delta x : \tau_1 \vdash_\Delta m : \tau_2 \# \Sigma \]

(S:Bnd:Term)
\[ \Delta; \Gamma \vdash_\Delta t : \tau_1 \# S \]
\[ \Delta, a : \tau_1; \Gamma \vdash_\Delta m : \tau_2 \# \{ a : S, a' : S' \} \]

\[ \Gamma \vdash_\Delta F : \langle \tau_1 \rangle (\tau_2) \leadsto \tau_3 \# F \]

(S:Def)
\[ f : \langle \tau_1 \rangle (\tau_2) \leadsto \tau_3 \# F \]
\[ \text{typedef } F = \forall a : \tau_1. \Sigma \]
\[ a : \tau_1; x : \tau_2 \Gamma \vdash_\Delta m : \tau_3 \# \Sigma \]

\[ \Gamma \vdash_\Delta \text{def } f(a)(x) = m : \langle \tau_1 \rangle (\tau_2) \leadsto \tau_3 \# F \]

Figure 25. Typing rules for terms, commands, and function definitions.

\[ s \vdash_\Sigma S \downarrow \sigma \]
\[ \sigma \vdash_\sigma \Sigma \downarrow \sigma \]

(W:事件:原子)
\[ \text{atom } v : \tau \]

(W:事件:迹)
\[ \text{trace } \sigma \downarrow \Sigma \downarrow \sigma \]

(W:事件:分支)
\[ \text{inj } \sigma \downarrow \Sigma \downarrow \text{true } \downarrow \sigma \downarrow \text{false } \downarrow \sigma \]

(W:事件:调用)
\[ \text{typedef } F = \forall a : \tau. \Sigma \]
\[ e \uparrow_\sigma \Sigma \downarrow \{ v_a / a \} \downarrow \sigma \downarrow \sigma \]
\[ \text{fold } \sigma \vdash_\Sigma F(e) \downarrow \sigma \]

(W:事件:结果)
\[ e \uparrow_\sigma \Sigma \downarrow \{ v_1 / a \} \downarrow \sigma \]

(W:事件:绑定)
\[ s \vdash_\Sigma S \downarrow \sigma \]
\[ \{ a : S, a' : S' \} \downarrow \sigma \]

Figure 26. Typing rules for traces and events.
### A.2 Measure semantics

Before developing the measure semantics for FCC programs, we first review the basics of measure theory briefly.

**σ-algebras** A σ-algebra on a set $X$ is a collection $\Phi$ of subsets of $X$ that includes $X$ itself and is closed under complement and under countable unions. Let $F$ be an arbitrary family of subsets of $X$, then there exists a unique smallest σ-algebra that is a superset of $F$. This σ-algebra is said to be generated by $F$, denoted by $\sigma(F)$.

**Measurable spaces** A measurable space $(X, \Phi)$ consists of an underlying set and its σ-algebra.

**Extended nonnegative reals** $\mathbb{R}^+$ is the set of nonnegative reals extended with infinity $\infty$.

**Measures** Let $(X, \Phi)$ be a measurable space. A function $\mu : \Phi \to \mathbb{R}_+$ is called a measure if it satisfies nonnegativity, null empty set, and countable additivity.

**Measure spaces** The triple $(X, \Phi, \mu)$ is called a measure space.

**σ-finite measures** The measure $\mu$ of a measure space $(X, \Phi, \mu)$ is σ-finite if $X$ can be decomposed into a countable union of sets with finite measure.

**Lebesgue measure** The Lebesgue measure is a σ-finite measure for $n$-dimensional Euclidean space, corresponding to length, area and volume for $n = 1, 2, 3$.

**Borel spaces** A set is a Borel set if and only if it is obtained by countable union, countable intersection, and relative complement from open sets in a topological space. For a topological space $(X, \tau)$, the collection of all Borel sets on $X$ forms a σ-algebra, known as Borel σ-algebra, notated $\mathcal{B}(X)$. A measurable space $(X, \Phi)$ is said to be a standard Borel space, if $\Phi$ is a Borel σ-algebra generated by a complete metric space on $X$. A measurable space $(X, \Phi)$ is a standard Borel space if and only if it is isomorphic to $\mathbb{R}_+$ or a subset of $\mathbb{N}$.

**Product measure space** Given two measure spaces $(X_1, \Phi_1, \mu_1)$ and $(X_2, \Phi_2, \mu_2)$, the product measurable space is defined as

$$
(X_1, \Phi_1) \otimes (X_2, \Phi_2) \overset{\text{def}}{=} (X_1 \times X_2, \sigma(\{B_1 \times B_2 \mid B_1 \in \Phi_1 \text{ and } B_2 \in \Phi_2\})).
$$

The product measure is defined as $(\mu_1 \otimes \mu_2)(B_1 \times B_2) \overset{\text{def}}{=} \mu_1(B_1) \cdot \mu_2(B_2)$. The measure $\mu_1 \otimes \mu_2$ is σ-finite if both $\mu_1$ and $\mu_2$ are σ-finite.

**Coprodut measure space** Given two measure spaces $(X_1, \Phi_1, \mu_1)$ and $(X_2, \Phi_2, \mu_2)$, the coproduct measure space is defined as

$$
(X_1, \Phi_1) \coprod (X_2, \Phi_2) \overset{\text{def}}{=} (X_1 \uplus X_2, \sigma(\{B_1 \uplus B_2 \mid B_1 \in \Phi_1 \text{ and } B_2 \in \Phi_2\})).
$$

The coproduct measure is defined as $(\mu_1 \uplus \mu_2)(B_1 \uplus B_2) \overset{\text{def}}{=} \mu_1(B_1) + \mu_2(B_2)$. Countable coproduct preserves σ-finiteness.

**Measurable spaces denoted by types.** The definition of $\llbracket \tau \rrbracket$ for each data type $\tau$ as a standard Borel space is standard:

- $\llbracket \text{unit} \rrbracket \overset{\text{def}}{=} (\{\text{unit}\}, \wp(\{\text{unit}\}))$
- $\llbracket \text{false} \rrbracket \overset{\text{def}}{=} (\{\text{false}\}, \wp(\{\text{false}\}))$
- $\llbracket \mathbb{R}_{[0,1]} \rrbracket \overset{\text{def}}{=} ((0, 1), \mathcal{B}(0, 1))$
- $\llbracket \mathbb{R}_+ \rrbracket \overset{\text{def}}{=} ((0, \infty), \mathcal{B}(0, \infty))$
- $\llbracket \mathbb{R} \rrbracket \overset{\text{def}}{=} ((-\infty, \infty), \mathcal{B}(-\infty, \infty))$
- $\llbracket \mathbb{N}_{\leq n} \rrbracket \overset{\text{def}}{=} \{0, 1, \cdots, n-1\}, \wp(\{0, 1, \cdots, n-1\})$
We can now define the step-indexed measurable spaces $\Sigma$ (resp. event type $S$), we construct a standard Borel space $\Sigma$ (resp. $S$). To handle recursive trace types, we first define a step-indexed version $\Sigma^n$ (resp. $S^n$). The definition uses helper injections that map sets formed by the substructures of traces and events into the corresponding sets of traces and events:

$$I_{a_1, \ldots, a_N} (A) \defeq \{ \{ a_1 : s_1, \ldots, a_N : s_N \} \mid (s_1, \ldots, s_N) \in A \}$$

where $a_1, \ldots, a_N$ are lexicographically ordered. The lexicographical ordering is on variable names. A canonical ordering of the term variables is needed, because traces are unordered maps but the construction of $N$-tuples requires an ordering.

We can now define the step-indexed measurable spaces $\Sigma^n$ and $S^n$, by double induction, first on the step index, and then on the structure of types:

$$\{(a_1 : s_1, \ldots, a_N : s_N)_e\}^n \defeq \left( I_{a_1, \ldots, a_N} (X), \{ I_{a_1, \ldots, a_N} (B) \mid B \in \Phi \} \right)$$

where $(X, \Phi) \defeq \bigotimes_{i=1}^N [S_i]^n$ and $(a_i)_{i=1}^N$ is the lexicographical reordering of $(a_i)_{i=1}^N$

$$n$$

where $(X, \Phi) \defeq [r]$ and typedef $F = \forall a : r. \Sigma$

Measurable spaces $\Sigma$ and $S$ can then be defined by induction on the structure of types, with the $[F(e)]$ case taking the fixed point:

$$\{(a_1 : s_1, \ldots, a_N : s_N)_e\} \defeq \left( I_{a_1, \ldots, a_N} (X), \{ I_{a_1, \ldots, a_N} (B) \mid B \in \Phi \} \right)$$

where $(X, \Phi) \defeq \bigotimes_{i=1}^N [S_i]$ and $(a_i)_{i=1}^N$ is the lexicographical reordering of $(a_i)_{i=1}^N$

$$n$$

where $(X, \Phi) \defeq [r]$ and typedef $F = \forall a : r. \Sigma$

$$n$$

where $(X, \Phi) \defeq \bigotimes_{i=1}^N [S_i]$ and $(a_i)_{i=1}^N$ is the lexicographical reordering of $(a_i)_{i=1}^N$

$$n$$

where $(X, \Phi) \defeq [r]$ and typedef $F = \forall a : r. \Sigma$

$$n$$

where $(X, \Phi) \defeq \bigotimes_{i=1}^N [S_i]$ and $(a_i)_{i=1}^N$ is the lexicographical reordering of $(a_i)_{i=1}^N$

$$n$$

where $(X, \Phi) \defeq [r]$ and typedef $F = \forall a : r. \Sigma$

$$n$$

where $(X, \Phi) \defeq \bigotimes_{i=1}^N [S_i]$ and $(a_i)_{i=1}^N$ is the lexicographical reordering of $(a_i)_{i=1}^N$

$$n$$

where $(X, \Phi) \defeq [r]$ and typedef $F = \forall a : r. \Sigma$

$$n$$

where $(X, \Phi) \defeq \bigotimes_{i=1}^N [S_i]$ and $(a_i)_{i=1}^N$ is the lexicographical reordering of $(a_i)_{i=1}^N$

$$n$$

where $(X, \Phi) \defeq [r]$ and typedef $F = \forall a : r. \Sigma$

$$n$$

where $(X, \Phi) \defeq \bigotimes_{i=1}^N [S_i]$ and $(a_i)_{i=1}^N$ is the lexicographical reordering of $(a_i)_{i=1}^N$

$$n$$

where $(X, \Phi) \defeq [r]$ and typedef $F = \forall a : r. \Sigma$

$$n$$

where $(X, \Phi) \defeq \bigotimes_{i=1}^N [S_i]$ and $(a_i)_{i=1}^N$ is the lexicographical reordering of $(a_i)_{i=1}^N$

$$n$$

where $(X, \Phi) \defeq [r]$ and typedef $F = \forall a : r. \Sigma$

$$n$$

where $(X, \Phi) \defeq \bigotimes_{i=1}^N [S_i]$ and $(a_i)_{i=1}^N$ is the lexicographical reordering of $(a_i)_{i=1}^N$

$$n$$

where $(X, \Phi) \defeq [r]$ and typedef $F = \forall a : r. \Sigma$

$$n$$

where $(X, \Phi) \defeq \bigotimes_{i=1}^N [S_i]$ and $(a_i)_{i=1}^N$ is the lexicographical reordering of $(a_i)_{i=1}^N$

$$n$$

where $(X, \Phi) \defeq [r]$ and typedef $F = \forall a : r. \Sigma$

$$n$$

where $(X, \Phi) \defeq \bigotimes_{i=1}^N [S_i]$ and $(a_i)_{i=1}^N$ is the lexicographical reordering of $(a_i)_{i=1}^N$

$$n$$

where $(X, \Phi) \defeq [r]$ and typedef $F = \forall a : r. \Sigma$

$$n$$

where $(X, \Phi) \defeq \bigotimes_{i=1}^N [S_i]$ and $(a_i)_{i=1}^N$ is the lexicographical reordering of $(a_i)_{i=1}^N$

$$n$$

where $(X, \Phi) \defeq [r]$ and typedef $F = \forall a : r. \Sigma$

$$n$$

where $(X, \Phi) \defeq \bigotimes_{i=1}^N [S_i]$ and $(a_i)_{i=1}^N$ is the lexicographical reordering of $(a_i)_{i=1}^N$
**Stock measures.** The stock measure $\mu_\tau$ on the measurable space $[\tau]$ denoted by a data type $\tau$ is standard, with $\mu_\tau$, $\mu_\Sigma$, and $\mu_{\Sigma_1}$ being the counting measures on their respective spaces, and $\mu_{\Sigma_1}$, $\mu_{\Sigma_2}$, and $\mu_{\Sigma}$ the Lebesgue measure.

For each trace type $\Sigma$ (resp. event type $S$), we define a stock measure $\mu_{\Sigma}$ (resp. $\mu_S$) on its semantic domain $[\Sigma]$ (resp. $[S]$). Again, we first define a step-indexed version, $\mu^n_\Sigma$ and $\mu^n_S$, via double induction on the index and the structure of types:

$$\mu^n_{\{a_1,S_1,\ldots,a_N:S_N\}}\left(I_{a_1,\ldots,a_N}(B)\right) \overset{\text{def}}{=} \left(\bigotimes_{i=1}^N \mu^n_{S_i}\right)(B) \quad \text{where } \{a_i\}_{i=1}^N \text{ is the lexicographical reordering of } (a_i)_{i=1}^N$$

$$\mu^n_{\text{atom } \tau}(I_{\text{atom }}(B)) \overset{\text{def}}{=} \mu_\tau(B)$$

$$\mu^n_{\text{trace } \Sigma}(I_{\text{trace }}(B)) \overset{\text{def}}{=} \mu^n_{\Sigma}(B)$$

$$\mu^n_{\Sigma_1,\Sigma_2}(I_{\text{inj }}(B)) \overset{\text{def}}{=} \left(\mu^n_{\Sigma_1} \uplus \mu^n_{\Sigma_2}\right)(B)$$

$$\mu^n_{\text{fold } (e)}(B) \overset{\text{def}}{=} \left(\bigotimes_{i=0}^n \mu^n_{e(i/a)}\right)(B) \quad \text{where } \text{typedef } F = \forall a : \tau, \Sigma$$

Measures $\mu^n_{\Sigma}$ and $\mu^n_S$ are then defined by induction on the structure of types:

$$\mu_{\{a_1,S_1,\ldots,a_N:S_N\}}\left(I_{a_1,\ldots,a_N}(B)\right) \overset{\text{def}}{=} \left(\bigotimes_{i=1}^N \mu_{S_i}\right)(B) \quad \text{where } \{a_i\}_{i=1}^N \text{ is the lexicographical reordering of } (a_i)_{i=1}^N$$

$$\mu_{\text{atom } \tau}(I_{\text{atom }}(B)) \overset{\text{def}}{=} \mu_\tau(B)$$

$$\mu_{\text{trace } \Sigma}(I_{\text{trace }}(B)) \overset{\text{def}}{=} \mu_\Sigma(B)$$

$$\mu_{\Sigma_1,\Sigma_2}(I_{\text{inj }}(B)) \overset{\text{def}}{=} \left(\mu_{\Sigma_1} \uplus \mu_{\Sigma_2}\right)(B)$$

$$\mu_{\text{fold } (e)}(B) \overset{\text{def}}{=} \left(\bigotimes_{i=0}^n \mu_{e(i/a)}\right)(B)$$

The measures are $\sigma$-finite, because the measures $\mu_\tau$ are $\sigma$-finite, and because $\sigma$-finiteness is preserved by finite products and countable coproducts, which are used to construct $\mu_{\Sigma}$ and $\mu_S$.

**Definition A.1 (Restatement of Definition 4.1).** A closed command $m$ of trace type $\Sigma (\bullet; \bullet \leftarrow_m m : \tau \# \Sigma)$ induces a measure $[\sigma]$ over a measurable space $[\Sigma]$ of traces:

$$P_m(\sigma) \overset{\text{def}}{=} \begin{cases} 1, & \text{if } \sigma \vdash_m m \leadsto_{\sigma} \omega \\ 0, & \text{otherwise} \end{cases}$$

$$[m](A) \overset{\text{def}}{=} \int_A P_m(\sigma) \mu_{\Sigma}(d\sigma)$$
B METATHEORETICAL DEVELOPMENT FOR FCC

B.1 Theorems in Section 4

PROPOSITION B.1. If $d : \text{dist}(\tau)$ and $v$ is a value, then $v : \tau$ if and only if $v \in d.\text{support}$ (i.e., $d.\text{density}(v) > 0$).

PROOF. Appeal to mathematical properties of primitive distributions.

PROPOSITION B.2 (Preservation and strong normalization of STLC).

- If $\Delta ; \Gamma \vdash e : \tau$, and $e \downarrow v$, then $\Delta ; \Gamma \vdash v : \tau$.
- If $\Delta ; \Gamma \vdash e : \tau$, then there exists a value $v$ such that $e \downarrow v$.

PROOF. Appeal to type soundness and strong normalization of the simply-typed lambda calculus.

LEMMA B.3.

(1) If $\sigma \vdash_m m \|_w v$, then $w > 0$.
(2) If $s \vdash_t t \|_w v$, then $w > 0$.

PROOF. By mutual structural induction on evaluation derivations.

LEMMA B.4 (Substitution of Expressions).

(1) If $\Delta ; \Gamma, x : t_1 \vdash e : t_2 \# S$ and $\Delta$; $\Gamma \vdash v : t_1$, then $\Delta$; $\Gamma \vdash t \{v/x\} : t_2 \# S$.
(2) If $\Delta, \psi : t_1 ; \Gamma \vdash t : t_2 \# S$ and $\Delta$; $\Gamma \vdash v : t_1$, then $\Delta$; $\Gamma \vdash t \{v/\psi\} : t_2 \# S \{\psi/a\}$.

PROOF. Appeals to the substitution lemma of the simply-typed lambda calculus.

LEMMA B.5 (Substitution for Commands).

(1) If $\Delta$; $\Gamma$; $x : t_1 \vdash e : t_2 \# \Sigma$ and $\Delta$; $\Gamma \vdash v : t_1$, then $\Delta$; $\Gamma \vdash x \{v/x\} : t_2 \# \Sigma$.
(2) If $\Delta, \psi : t_1 ; \Gamma \vdash e : t_2$ and $\Delta, \psi : t_1 \vdash v : t_1$, then $\Delta, \psi : t_1 \vdash e \{v/\psi\} : t_2$.

PROOF. We show several nontrivial cases for (item 4) and (item 2).

Proof by mutual induction.

Case: (S-BND-EXPR)

(1) $\Delta, \alpha : t_1 ; \Gamma \vdash_m \text{ret}(e) : t_2 \# \{e\} \Delta$ and $\Delta ; \Gamma \vdash v : t_1$ (ASSUMPTION)
(2) $\Delta, \alpha : t_1 ; \Gamma \vdash e : t_2$ (INVERSION) $\&$ (1)
(3) $\Delta ; \bullet \vdash \{e\} : t_2$ (LEMMA B.4) $\&$ (1) $\&$ (2)
(4) $\Delta ; \Gamma \vdash_m \text{ret}(e \{v/\alpha\}) : t_2 \# \{e\{v/\alpha\}\}$ S-BND-EXPR $\&$ (3) $\&$ (5)

Case: (S-BND-TERM)

(1) $\Delta, \alpha : t_1 ; \Gamma \vdash_m b = t : m : t_2 \# \Sigma$ and $\Delta$; $\Gamma \vdash v : t_0$ (ASSUMPTION)
(2) $\Delta, \alpha : t_1 ; \Gamma \vdash e : t_2 \# S$ and $\Delta, \alpha : t_1 ; \Gamma \vdash b : \tau_b$, $\Gamma \vdash_m m : t_2 \# \{\tau_b\} S \{\alpha\}$ (INVERSION) $\&$ (1)
(3) $\Delta, \alpha : t_1 ; \Gamma \vdash t \{v/\alpha\} : t_2 \# S \{v/\alpha\}$ (I.H.) $\&$ (1) $\&$ (2)
(4) $\Delta, \alpha : t_1 ; \Gamma \vdash v : t_1$ (WEAKENING) $\&$ (1)
Theorem B.6 (Restatement of Theorem 4.3).

(1) If \(\bullet \vdash m : \tau \# \Sigma\), and \(\sigma \vdash m \downarrow^w v\), then \(\bullet \vdash \rho \vdash \tau \# \Sigma \downarrow^w v\).

(2) If \(\bullet \vdash \rho \vdash \tau \# S\) and \(s \vdash t \downarrow^w v\), then \(\bullet \vdash \rho \vdash \tau \# S \downarrow^w v\).

Proof. By mutual induction on the derivation of \(\sigma \vdash m \downarrow^w v\), and \(s \vdash t \downarrow^w v\) followed by inversion on \(\bullet \vdash m : \tau \# \Sigma\) and \(\bullet \vdash \rho \vdash \tau \# S\) respectively. We show several nontrivial cases; others are similar to one of these cases.

Case: (E:RET)

\[
\begin{align*}
\text{(E:RET)} & \quad e \downarrow^w v \\
\{ \} \vdash m \text{ ret}(e) \downarrow^1 v \\
\Delta; \bullet \vdash e : \tau \\
\Delta; \bullet \vdash \rho \vdash \tau \# \{ e \} \\
\text{(Assumption)} & \quad \text{(Proposition B.2) (1)} \quad \text{W:Trace:RET (1)}
\end{align*}
\]
\[ \begin{aligned}
\text{Case: (E:Bnd:Expr)} & \quad \text{(E:Bnd:Expr)} \\
\quad e \downarrow v_e & \quad \sigma \vdash_m m \{v_e/x\} \downarrow^w v \\
\quad \sigma \vdash_m x = e; m \downarrow^w v & \\
(1) & \quad \bullet; \bullet \vdash_e x = x_1 \quad \text{(Assumption)} \\
(2) & \quad \bullet; \bullet \vdash_e x = x_1 \\
(3) & \quad \bullet; x : x_1 \vdash_m m : x_2 \quad \text{(Assumption)} \\
(4) & \quad \bullet; \bullet \vdash_m m \{v_e/x\} : x_2 \quad \text{(Assumption)} \\
(5) & \quad \bullet; \bullet \vdash_m m \{v_e/x\} \downarrow^w v \\
(6) & \quad \bullet; \bullet \vdash_v v = v_2 \\
(7) & \quad \sigma \vdash \Sigma \downarrow v \\
\text{(S:Bnd:Expr)} & \quad \Delta; \Gamma \vdash_e e : x_1 \\
\Delta; \sigma \vdash \Delta_m x = e; m : x_2 \quad \text{(Lemma B.5)} \quad (1) \\
\Delta; \Gamma \vdash_m m : x_2 \quad \text{(Lemma B.5)} \quad (2) \quad (3) \\
\text{(Case: (E:Bnd:Term))} & \quad \text{(E:Bnd:Term)} \\
\quad s \vdash t \quad \downarrow^w v_1 & \quad \begin{aligned}
\begin{bmatrix}
\alpha' : S' \\
\vdash_m m \{v_1/a\} \downarrow^w v_2
\end{bmatrix}
\end{aligned} \\
\left( a : S, \alpha' : S' \right) \vdash_m a = t; m \downarrow^w v_2 \\
\Delta; \Gamma \vdash a = t; m : x_2 \quad \text{(Lemma B.5)} \quad (2) \quad (3) \\
\Delta; \Gamma \vdash_m a = t; m : x_2 \quad \text{(Lemma B.5)} \quad (1) \\
\text{(S:Bnd:Term)} & \quad \Delta; \Gamma \vdash t : x_1 \neq S \\
\Delta, a : t_1; \Gamma \vdash_m m : \Sigma \quad \text{(Lemma B.5)} \quad (2) \quad (3) \\
\Delta; \Gamma \vdash_m m : \Sigma \quad \text{(Lemma B.5)} \quad (1) \\
\text{(Case: (E:Observe))} & \quad \text{(E:Observe)} \\
\quad v \in d.\text{support} \quad d \equiv \text{d.\text{density}(v_2)} & \quad \sigma \vdash_m m \downarrow^w v \\
\sigma \vdash_m \text{observe}(e_1; e_2); m \downarrow^w v & \\
\left( 1 \right) & \quad \bullet; \bullet \vdash_m m : \Sigma \quad \text{(Assumption)} \\
\left( 2 \right) & \quad \bullet; \bullet \vdash_v v = v_2 \\
\text{(S:Observe)} & \quad \Delta; \Gamma \vdash_e e_1 : \text{dist}(r) \\
\Delta; \Gamma \vdash_m m : \Sigma \quad \text{(Assumption)} \\
\Delta; \Gamma \vdash_m \text{observe}(e_1; e_2); m : \Sigma \\
\Delta; \Gamma \vdash_m m : \Sigma \quad \text{(I.H.)} \quad (1) \\
\text{(Case: (E:Cmd))} & \quad \text{(E:Cmd)} \\
\quad \sigma \vdash_m m \downarrow^w v & \quad \text{(S:Cmd)} \\
\quad \vdash \sigma \vdash_m m \downarrow^w v & \quad \Delta; \Gamma \vdash_m m : \Sigma \quad \text{(Assumption)} \\
\text{(1)} & \quad \bullet; \bullet \vdash_m m : \Sigma \quad \text{(Assumption)} \\
\text{(2)} & \quad \bullet; \bullet \vdash_v v = v_2 \\
\text{(3)} & \quad \text{trace} \sigma \vdash \Sigma \downarrow v \\
\text{(W:Event:Trace)} & \quad \text{(W:Event:Trace)} \\
\end{aligned} \]
Theorem B.7 (Restatement of Theorem 4.4).

1. If \( \bullet \cdot \vdash m : \tau \# \Sigma \) and \( \sigma \vdash \Sigma \| v \), then there exist \( w \) such that \( \sigma \vdash m \| w \| v \).
2. If \( \bullet \cdot \vdash t : \tau \# S \) and \( s \vdash S \| v \), then there exist \( w \) such that \( s \vdash t \| w \| v \).
PROOF. By mutual induction on the derivation of \( \bullet \cdot t_m : \tau \# \Sigma \), and \( \bullet \cdot t_I : \tau \# S \), followed by nested induction on \( \sigma \vdash _\tau \Sigma \parallel \nu \) and \( s \vdash _\tau S \parallel \nu \) respectively.

**Case:** (S:RETURN)

\[
\begin{align*}
\Delta; & \cdot \vdash e : \tau \\
\Delta; & \Gamma \vdash _\tau \text{ret}(e) : \tau \# \{}e\\}
\end{align*}
\]

1. \( \bullet \cdot \vdash e : \tau \) and \( e \parallel \nu \)
2. \( \{}e \vdash \text{ret}(e) \parallel ^{1} \nu \)

(ASSUMPTION)

E:RET \& (1)

**Case:** (S:BND:EXPR)

\[
\begin{align*}
\Delta; & \cdot \vdash e : \tau_1 \\
\Delta; & \Gamma, x : \tau_1 \vdash _\tau m : \tau_2 \# \Sigma \\
\Delta; & \Gamma \vdash _\tau x = e ; m : \tau_2 \# \Sigma
\end{align*}
\]

1. \( \bullet \cdot \vdash e : \tau_1 \) (ASSUMPTION)
2. There exists a value \( \nu e \) s.t. \( e \parallel \nu e \) and \( \bullet \cdot \vdash \nu e : \tau_1 \) (PROPOSITION B.2) \& (1)
3. \( \bullet \cdot x : \tau_1 \vdash _\tau m : \tau_2 \# \Sigma \) (ASSUMPTION)
4. \( \bullet \cdot \vdash m \{\nu e / x\} : \tau_2 \# \Sigma \) (LEMMA B.5) \& (2)
5. \( \sigma \vdash _\tau \Sigma \parallel \nu \) (ASSUMPTION)
6. There exists \( w \) and \( v \) s.t. \( \sigma \vdash _\tau m \{\nu e / x\} \parallel ^{w} \nu \) (I.H.) \& (4) \& (5)
7. \( \sigma \vdash _\tau x = e ; m \parallel ^{w} \nu \) E:BND:EXPR \& (2) \& (6)

**Case:** (S:BND:TERM)

\[
\begin{align*}
\Delta; & \cdot \vdash t : \tau_1 \# S \\
\Delta; & \cdot \vdash a : \tau_1 ; \Gamma \vdash _\tau m : \tau_2 \# \{a' : S'\}_e
\end{align*}
\]

\[
\begin{align*}
\Delta; & \Gamma \vdash _\tau a = t ; m : \tau_2 \# \{a : S, a' : S'\}_e \\
\end{align*}
\]

1. \( \bullet \cdot \vdash t : \tau_1 \# S \) and \( s \vdash _\sigma S \parallel \nu_1 \) (ASSUMPTION)
2. There exists \( w_1 \) s.t. \( s \vdash _\sigma \nu_1 \parallel ^{w_1} \nu_1 \) (I.H.) \& (1)
3. \( a : \tau_1 ; \cdot \vdash _\tau m : \tau_2 \# \{a' : S'\}_e \) (ASSUMPTION)
4. \( \bullet \cdot \vdash _\tau v_1 : \tau_1 \) (THEOREM B.6) \& (1) \& (2)
5. \( \bullet \cdot \vdash m \{\nu_1 / a\} : \tau_2 \# \{a' : S'\}_e \{\nu_1 / a\} \) (LEMMA B.5) \& (3) \& (4)
6. \( \{a : S, a' : S'\}_e \vdash _\sigma \{a : S, a' : S'\}_e \parallel \nu \) (ASSUMPTION)
7. \( \{a' : S'\}_e \vdash _\sigma \{a' : S'\}_e \{\nu_1 / a\} \parallel \nu \) W:TRACE:BND \& (1) \& (6)
8. There exists \( w_2 \) s.t. \( \{a' : S'\}_e \vdash _\sigma \{w_1 / a\} \parallel ^{w_2} \nu_2 \) (I.H.) \& (5) \& (7)
9. \( \{a : s, a' : S'\}_e \vdash _\sigma \{a : s, a' : S'\}_e \parallel \nu \) E:BND:TERM \& (2) \& (8)

**Case:** (S:OBSERVE)

\[
\begin{align*}
\Delta; & \cdot \vdash e_1 : \text{dist}(\tau) \\
\Delta; & \cdot \vdash e_2 : \tau \\
\Delta; & \cdot \vdash m : \tau \# \Sigma
\end{align*}
\]

\[
\begin{align*}
\Delta; & \Gamma \vdash _\sigma \text{observe}(e_1 ; e_2) ; m : \tau \# \Sigma \\
\end{align*}
\]

1. \( \bullet \cdot \vdash m : \tau \# \Sigma \) and \( \sigma \vdash _\tau \Sigma \parallel \nu \) (ASSUMPTION)
(2) there exists \( w_2 \) s.t. \( \sigma \vdash m \parallel^{w_2} v \)  
(I.H.) \& (1)  
(3) \( \bullet \vdash e_1 : \text{dist}(\tau) \) and \( \bullet \vdash e_2 : \tau \)  
(Assumption)  
(4) \( e_1 \parallel d, e_2 \parallel v \) and \( \bullet \vdash d : \text{dist}(\tau) \) and \( \bullet \vdash \tau : \tau \)  
(Proposition B.2) \& (3)  
(5) \( \sigma \vdash \text{observe}(e_1; e_2) ; m \parallel^{w_1 \cdot w_2} v \)  
(Proposition B.1) \& (4)  
(6) \( \sigma \vdash m \)  
E:Observe \& (2) \& (4) \& (5)  

Case: (S:Cmd)  

\[
\begin{align*}
\Delta; \Gamma & \vdash m : \tau \neq \Sigma \\
\Delta; \Gamma & \vdash \text{sample}(e) : \tau \neq \text{atom} \tau \\
\end{align*}
\]

(1) \( \bullet \vdash e : \text{dist}(\tau) \)  
(Assumption)  
(2) \( e \parallel d \)  
(Proposition B.2) \& (1)  
(3) \( \bullet \vdash e : \text{dist}(\tau) \)  
(Theorem 4.3) \& (1) \& (2)  
(4) \( \bullet \vdash \tau : \tau \)  
(Assumption)  
(5) \( v \in d.\text{support} \) and \( w = d.\text{density}(v) \) and \( w > 0 \)  
(Proposition B.1) \& (3) \& (4)  
(6) \( \atom v \vdash \text{sample}(e) \parallel^w v \)  
E:Sample \& (2) \& (5)  

Case: (S:Branch)  

\[
\begin{align*}
\Delta; \bullet & \vdash e : 2 \\
\Delta; \Gamma & \vdash m_1 : \tau \neq \Sigma_1 \\
\Delta; \Gamma & \vdash m_2 : \tau \neq \Sigma_2 \\
\Delta; \Gamma & \vdash \text{ite}(e; m_1; m_2) : \tau \neq \Sigma_1 + \epsilon \Sigma_2 \\
\end{align*}
\]

(1) \( \bullet \vdash e : \text{dist}(\tau) \)  
(Assumption)  
(2) \( e \parallel v_b \)  
(Assumption)  
(3) \( \there exists w s.t. \sigma \vdash m_1 \parallel^w v \)  
(Assumption)  
(4) \( \text{inj} \sigma_3 \Sigma_\text{false} \parallel^t \text{false} \)  
(I.H.) \& (1)  
E:Branch \& (2) \& (3)  

Case: (S:Call)  

\[
\begin{align*}
\Delta; \Gamma & \vdash \text{call}(f; e_1; e_2) : \tau_3 \neq F(e_1) \\
\end{align*}
\]

(1) \( \bullet \vdash e_2 : \tau_2 \)  
(Assumption)  
(2) \( \there exists \text{a value } v_x \text{ s.t. } e_2 \parallel v_x \)  
(Proposition B.2) \& (1)  
(3) \( \bullet \vdash e_1 : \tau_1 \) and \( e_1 \parallel v_a \)  
(Assumption)  
(4) \( \bullet \vdash v_a : \tau_1 \) and \( \bullet \vdash v_x : \tau_2 \)  
(Proposition B.2) \& (1) \& (2) \& (3)  
(5) \( a : \tau_1 ; \text{Lemma B.5} \& (4) \) \& (4) \& (7)  
\( \text{Lemma B.5} \& (4) \) \& (4) \& (7)  
(6) \( \bullet \vdash m \{ v_a / a \} \parallel v_x \parallel \Sigma \neq v_x / x \)  
(7) \( \sigma \parallel_\delta \Sigma \{ v/a \} \parallel v \)  
(Assumption)  
(8) \( \there exists w s.t. \sigma \vdash m \{ v_a / a \} \parallel^w v \)  
(I.H.) \& (6) \& (7)  
(9) \( \text{typedef } F = \forall a : \tau_1. \Sigma \)  
(Assumption)
We show the most interesting case, namely the translation of a command.

\[ \text{CPFV}_{m}(\cdot) \]

\[ \Delta \vdash_{\Sigma} \text{call}(f; e_1; e_2) :: \nu \]

**LEMMA B.8.** If \( \bullet ; \bullet \vdash_{m} 1 : \tau \# \Sigma \) and \( \bullet ; \vdash_{m} 2 : \tau \# \Sigma \), then it holds that \( P_{\text{m}_{1}}(\sigma) \neq 0 \) if and only if \( P_{\text{m}_{2}}(\sigma) \neq 0 \).

**Proof.** The "only if" direction mirrors the "if" direction. Below we show the "if" direction.

1. \( P_{\text{m}_{1}}(\sigma) > 0 \)  
   Assumption
2. \( \bullet ; \vdash_{m} 1 : \tau \# \Sigma \)  
   Assumption
3. \( \bullet ; \vdash_{m} 2 : \tau \# \Sigma \)  
   Assumption
4. \( \sigma \vdash_{m} 1 \downarrow \nu \)  
   Theorem B.6 \( \sigma \downarrow \nu \)
5. \( \sigma \vdash_{m_{2}} \downarrow \nu \)  
   Theorem B.7 \( \sigma \downarrow \nu \)
6. \( P_{\text{m}_{2}}(\sigma) = w_{2} > 0 \)  
   Definition A.1 \& Lemma B.3 \( \sigma \downarrow \nu \)

**THEOREM B.9 (RESTATEMENT OF THEOREM 4.5).** If \( \bullet ; \bullet \vdash_{m} 1 : \tau \# \Sigma \) and \( \bullet ; \vdash_{m} 2 : \tau \# \Sigma \), then for any measurable set \( A \) of traces, \( \mu(A) \neq 0 \) if and only if \( \mu(A) \neq 0 \).

**Proof.** The proof uses the fact that a Lebesgue integral is zero if and only if the integrand is zero almost everywhere.

\[
\begin{align*}
\mu(2)(A) &= \int_{E} P_{\text{m}_{2}}(\sigma) d\sigma = 0 \\
\iff &\lambda_{\sigma}\{\sigma | \sigma \vdash P_{\text{m}_{2}}(\sigma) \cdot 1_{E}(\sigma) > 0\} = 0 \\
\iff &\lambda_{\sigma}\{\sigma | \sigma \vdash P_{\text{m}_{1}}(\sigma) \cdot 1_{E}(\sigma) > 0\} = 0 \\
\iff &\mu(1)(A) = \int_{E} P_{\text{m}_{1}}(\sigma) d\sigma = 0
\end{align*}
\]

\[ \Box \]

### B.2 Theorems in Section 6

Define functions \( \text{CPFV}_{m}(\cdot) \) and \( \text{CPFV}_{l}(\cdot) \) that compute free term variables in checkpoints within a command or a term:

\[
\begin{align*}
\text{CPFV}_{m}(\text{ret}(e)) &\overset{\text{def}}{=} \text{FV}(e) \\
\text{CPFV}_{m}(x = e; m) &\overset{\text{def}}{=} \text{CPFV}_{m}(m) \\
\text{CPFV}_{m}(a \equiv t; m) &\overset{\text{def}}{=} \text{CPFV}_{m}(t) \cup \text{CPFV}_{m}(m) \\
\text{CPFV}_{m}(\text{observe}(e_1; e_2); m) &\overset{\text{def}}{=} \text{CPFV}_{m}(m) \\
\text{CPFV}_{l}(\text{sample}(e)) &\overset{\text{def}}{=} \emptyset \\
\text{CPFV}_{l}(\text{ite}(e; m_1; m_2)) &\overset{\text{def}}{=} \text{FV}(e) \cup \text{CPFV}_{m}(m_1) \cup \text{CPFV}_{m}(m_2) \\
\text{CPFV}_{l}(\text{call}(f; e)) &\overset{\text{def}}{=} \text{FV}(e) \\
\text{CPFV}_{l}(m) &\overset{\text{def}}{=} \text{CPFV}_{m}(m)
\end{align*}
\]

**LEMMA B.10.**

1. If \( \mathcal{G}; C; \Delta' ; h : \mathbb{R} \vdash_{m} \langle \Delta; \Gamma \vdash_{m} m : \tau \# \Sigma \rangle \rightsquigarrow m' \) and \( \text{CPFV}_{m}(m) \subseteq \text{domain}(\Delta') \), then \( \Delta' : h : \mathbb{R} \vdash_{m} m' : \tau \# \Sigma \).
2. If \( \mathcal{G}; C; \Delta' ; h : \mathbb{R} \vdash_{l} \langle \Delta; \Gamma \vdash_{l} t : \tau \# S \rangle \rightsquigarrow t' \) and \( \text{CPFV}_{l}(t) \subseteq \text{domain}(\Delta') \), then \( \Delta' : h : \mathbb{R} \vdash_{l} t' : \tau \# S \).

**Proof.** By mutual structural induction on derivations of command translation and term translation. Below we show the most interesting case, namely the translation of a command.
Case: GG-CMD

\[ \text{TermBindings}(m) = (a_l = t_l)_{l=1}^{N}, \quad (a_l = t_l)_{l=1}^{N} \] reorders \( G \)

\( \forall i. \ G; C; \Delta', a_l : \tau_l, \ldots, a_{l_{i-1}} : \tau_{l_{i-1}}; h : \mathbb{R}; a_l \vdash (\Delta_l; \Delta_l' \vdash t_l; \tau_l \# S_l) \leadsto t'_l \)

\[ G; C; \Delta'; h : \mathbb{R} \vdash_m (\Delta; \Gamma \vdash_m \tau \# \{a_1 : S_1, \ldots, a_N : S_N\}_e) \leadsto a_l = t'_l; \ldots; a_{l_N} = t'_{l_N}; \text{ret}(e) \]

1. \( \Delta; \Gamma \vdash_m m : \tau \# \{a_1 : S_1, \ldots, a_N : S_N\}_e \)
2. \( \{a_l = t_l\}_{l=1}^{N} \) reorders \( G \)
3. \( \forall i. \Delta_i; \Gamma_i \vdash t_l : \tau_l \# S_l \)
4. \( \Delta, a_1 : \tau_1, \ldots, a_N : \tau_N; \bullet \vdash e : \tau \)
5. \( \Delta, a_l : \tau_l, \ldots, a_{l-1} : \tau_{l-1}; h : \mathbb{R} \vdash t'_l : \tau_l \# S_l \)
6. \( \forall i. \Delta'_l, a_l : \tau_l, \ldots, a_{l_{i-1}} : \tau_{l_{i-1}}; h : \mathbb{R} \vdash t'_l : \tau_l \# S_l \)
7. \( \text{CPFV}(t_l) \subseteq \text{domain}(\Delta') \cup \{a_l, \ldots, a_{l_{i-1}}\} \)
8. \( \forall i. \Delta'_l, a_l : \tau_l, \ldots, a_{l_{i-1}} : \tau_{l_{i-1}}; h : \mathbb{R} \vdash t'_l : \tau_l \# S_l \)
9. \( \tau \# \{a_1 : S_1, \ldots, a_N : S_N\}_e \)

\[ \text{inversion of (2)} \]

\( (i_l)_{l=1}^{N} \) is a permutation of \( (i)_{l=1}^{N} \)

\( \{a_l : S_l, \ldots, a_{l_N} : S_{l_N}\}_e = \{a_1 : S_1, \ldots, a_N : S_N\}_e \)

\( \Delta'; h : \mathbb{R} \vdash_m a_l = t'_l; \ldots; a_{l_N} = t'_{l_N}; \text{ret}(e) \)

\( (9) \vdash (11) \)

Theorem B.11 (Restatement of Theorem 6.1).

If \( \Gamma ; \{\text{ENTRY}\}; \bullet; \text{obs} : \mathbb{R} \vdash_m (\bullet; \text{obs} : \mathbb{R} \vdash_m m : \tau \# \Sigma) \leadsto m_g, \text{then } \bullet; \text{obs} : \mathbb{R} \vdash_m m_g : \tau \# \Sigma. \)

Proof. By Lemma B.10 and by command \( m \) being well-typed in an empty term-variable environment.

Theorem B.12 (Restatement of Theorem 6.2).

If \( \Gamma ; \{\text{ENTRY}\}; \bullet; \text{obs} : \mathbb{R} \vdash_m (\bullet; \text{obs} : \mathbb{R} \vdash_m m : \tau \# \Sigma) \leadsto m_g \) and \( \text{obs} : \mathbb{R} \), then for any measurable set \( A \) of traces, \( \lfloor m_g \{\text{obs}/\text{obs}\} \rfloor (A) \neq 0 \) if and only if \( \lfloor m_g \{\text{obs}/\text{obs}\} \rfloor (A) \neq 0. \)

Figure 27. Dataflow for a trace of the captcha model and two guides. In the model (a), RVs $c$ (captcha characters) are all correlated with each other. As a result, the model contains no conditional independence for Fidelio to exploit.
## D EXPERIMENTAL SETUPS AND SUPPLEMENTAL STATISTICS

<table>
<thead>
<tr>
<th>name</th>
<th>structure</th>
<th>#RV</th>
<th>#LOC (model)</th>
<th>#iteration</th>
<th>guide</th>
<th>capacity</th>
<th>validation loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>longrange (k=5)</td>
<td>loop</td>
<td>8</td>
<td>16</td>
<td>10000</td>
<td>Fidelia</td>
<td>1398</td>
<td>21.87529387</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LSTM</td>
<td>1382</td>
<td>24.49491011</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>mean-field</td>
<td>1386</td>
<td>25.38885483</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Fidelia</td>
<td>1398</td>
<td>40.37852367</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LSTM</td>
<td>1382</td>
<td>43.68544903</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>mean-field</td>
<td>1386</td>
<td>44.04598006</td>
</tr>
<tr>
<td>longrange (k=10)</td>
<td>loop</td>
<td>13</td>
<td>16</td>
<td>10000</td>
<td>Fidelia</td>
<td>1398</td>
<td>77.5934972</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LSTM</td>
<td>1382</td>
<td>81.31816754</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>mean-field</td>
<td>1386</td>
<td>81.24487837</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Fidelia</td>
<td>1398</td>
<td>152.1430431</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LSTM</td>
<td>1382</td>
<td>156.0312375</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>mean-field</td>
<td>1386</td>
<td>155.7906566</td>
</tr>
<tr>
<td>longrange (k=20)</td>
<td>loop</td>
<td>23</td>
<td>16</td>
<td>10000</td>
<td>sparse guide (Fidelia)</td>
<td>994</td>
<td>18.77669632</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>dense guide (Fidelia)</td>
<td>982</td>
<td>20.46032267</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>mean-field</td>
<td>1078</td>
<td>22.88487139</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Fidelia</td>
<td>1540</td>
<td>18.74158391</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LSTM</td>
<td>1533</td>
<td>18.99891774</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>mean-field</td>
<td>1588</td>
<td>20.48320966</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>sparse guide (Fidelia)</td>
<td>2023</td>
<td>18.72428359</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>dense guide (Fidelia)</td>
<td>2047</td>
<td>19.00601651</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LSTM</td>
<td>2116</td>
<td>19.97572875</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>mean-field</td>
<td>2128</td>
<td>22.75885654</td>
</tr>
<tr>
<td>longrange (k=40)</td>
<td>loop</td>
<td>43</td>
<td>16</td>
<td>10000</td>
<td>sparse guide (Fidelia)</td>
<td>5283</td>
<td>41.06646845</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>dense guide (Fidelia)</td>
<td>5215</td>
<td>43.5052426</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>mean-field</td>
<td>2196</td>
<td>45.76123151</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Fidelia</td>
<td>3642</td>
<td>41.04653035</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LSTM</td>
<td>3594</td>
<td>43.11532713</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>mean-field</td>
<td>3552</td>
<td>44.85451233</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>sparse guide (Fidelia)</td>
<td>5196</td>
<td>41.04030352</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>dense guide (Fidelia)</td>
<td>5298</td>
<td>43.11325543</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LSTM</td>
<td>5498</td>
<td>43.95748757</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>mean-field</td>
<td>5412</td>
<td>51.88407577</td>
</tr>
<tr>
<td>treebn (h=3)</td>
<td>BN</td>
<td>15</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>sparse guide (Fidelia)</td>
<td>2055</td>
<td>41.06646845</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>dense guide (Fidelia)</td>
<td>2154</td>
<td>43.5052426</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>mean-field</td>
<td>2196</td>
<td>45.76123151</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Fidelia</td>
<td>3642</td>
<td>41.04653035</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LSTM</td>
<td>3594</td>
<td>43.11532713</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>mean-field</td>
<td>3552</td>
<td>44.85451233</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>sparse guide (Fidelia)</td>
<td>5196</td>
<td>41.04030352</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>dense guide (Fidelia)</td>
<td>5298</td>
<td>43.11325543</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LSTM</td>
<td>5498</td>
<td>43.95748757</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>mean-field</td>
<td>5412</td>
<td>51.88407577</td>
</tr>
<tr>
<td>treebn (h=4)</td>
<td>BN</td>
<td>31</td>
<td>37</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>sparse guide (Fidelia)</td>
<td>2055</td>
<td>41.06646845</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>dense guide (Fidelia)</td>
<td>2154</td>
<td>43.5052426</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>mean-field</td>
<td>2196</td>
<td>45.76123151</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Fidelia</td>
<td>3642</td>
<td>41.04653035</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LSTM</td>
<td>3594</td>
<td>43.11532713</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>mean-field</td>
<td>3552</td>
<td>44.85451233</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>sparse guide (Fidelia)</td>
<td>5196</td>
<td>41.04030352</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>dense guide (Fidelia)</td>
<td>5298</td>
<td>43.11325543</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LSTM</td>
<td>5498</td>
<td>43.95748757</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>mean-field</td>
<td>5412</td>
<td>51.88407577</td>
</tr>
<tr>
<td>PCFG-Captcha</td>
<td>branching recursion</td>
<td>stochastic</td>
<td>237</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Fidelia</td>
<td>52 838</td>
<td>7.536001405</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LSTM</td>
<td>50 159</td>
<td>14.76856677</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>mean-field</td>
<td>52 105</td>
<td>10.29520308</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Fidelia</td>
<td>104 951</td>
<td>3.6027033</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LSTM</td>
<td>102 437</td>
<td>14.83619706</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>mean-field</td>
<td>100 217</td>
<td>7.961722895</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Fidelia</td>
<td>254 555</td>
<td>2.241801572</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LSTM</td>
<td>250 209</td>
<td>14.78410727</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>mean-field</td>
<td>250 687</td>
<td>8.149919943</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Fidelia</td>
<td>500 603</td>
<td>2.918338221</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LSTM</td>
<td>500 049</td>
<td>14.44708732</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>mean-field</td>
<td>501 999</td>
<td>10.2873026</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Fidelia</td>
<td>2 000 177</td>
<td>2.51201136</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LSTM</td>
<td>2 000 705</td>
<td>9.530934941</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>mean-field</td>
<td>2 000 321</td>
<td>8.379732159</td>
</tr>
<tr>
<td>name</td>
<td>structure</td>
<td>#RV</td>
<td>#LOC (model)</td>
<td>#iteration</td>
<td>guide</td>
<td>capacity</td>
<td>validation loss</td>
</tr>
<tr>
<td>-----------------</td>
<td>-----------</td>
<td>-----</td>
<td>--------------</td>
<td>------------</td>
<td>-------</td>
<td>----------</td>
<td>-----------------</td>
</tr>
<tr>
<td>Fidelia</td>
<td></td>
<td></td>
<td></td>
<td>2000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LSTM</td>
<td>1154</td>
<td>20.47216836</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>mean-field</td>
<td>1496</td>
<td>25.98369968</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fidelia</td>
<td>5469</td>
<td>13.36186348</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LSTM</td>
<td>5042</td>
<td>15.65118018</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>mean-field</td>
<td>5456</td>
<td>17.55795286</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fidelia</td>
<td>10505</td>
<td>25.98369968</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LSTM</td>
<td>10334</td>
<td>15.28437446</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>mean-field</td>
<td>10824</td>
<td>16.56617211</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fidelia</td>
<td>50591</td>
<td>11.82502852</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LSTM</td>
<td>50774</td>
<td>15.26491375</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>mean-field</td>
<td>52404</td>
<td>16.76144694</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fidelia</td>
<td>101481</td>
<td>11.93564912</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LSTM</td>
<td>10304</td>
<td>15.31487038</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>mean-field</td>
<td>10824</td>
<td>16.56617211</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ECOLI70</td>
<td>BN</td>
<td>46</td>
<td>54</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LSTM</td>
<td>152196</td>
<td>17.51372583</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>mean-field</td>
<td>152196</td>
<td>17.51372583</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fidelia</td>
<td>4416271</td>
<td>-1.21193114</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LSTM</td>
<td>4416081</td>
<td>-1.32465717</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>mean-field</td>
<td>4423701</td>
<td>1.64354298</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fidelia</td>
<td>25028</td>
<td>2.57001384</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LSTM</td>
<td>25006</td>
<td>2.57001384</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>mean-field</td>
<td>25006</td>
<td>2.57001384</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fidelia</td>
<td>50038</td>
<td>1.83395378</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LSTM</td>
<td>50022</td>
<td>2.47013403</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>mean-field</td>
<td>50022</td>
<td>2.47013403</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fidelia</td>
<td>74986</td>
<td>1.79265493</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LSTM</td>
<td>75035</td>
<td>2.35631491</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>mean-field</td>
<td>75035</td>
<td>2.35631491</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fidelia</td>
<td>100004</td>
<td>14.9902177</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LSTM</td>
<td>100166</td>
<td>16.42537214</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>mean-field</td>
<td>100166</td>
<td>16.42537214</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fidelia</td>
<td>153566</td>
<td>12.58098554</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LSTM</td>
<td>151052</td>
<td>15.17583296</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>mean-field</td>
<td>151052</td>
<td>15.17583296</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fidelia</td>
<td>20454</td>
<td>2.26109840</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LSTM</td>
<td>19794</td>
<td>257.8196954</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>mean-field</td>
<td>19794</td>
<td>257.8196954</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fidelia</td>
<td>49872</td>
<td>1.59185495</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LSTM</td>
<td>49782</td>
<td>152.3689971</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>mean-field</td>
<td>49782</td>
<td>152.3689971</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fidelia</td>
<td>80222</td>
<td>14.55081707</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LSTM</td>
<td>80184</td>
<td>152.3689971</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>mean-field</td>
<td>80184</td>
<td>152.3689971</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>gmm</td>
<td>branching recursion</td>
<td>116</td>
<td>49</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LSTM</td>
<td>20332</td>
<td>16.21922317</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>mean-field</td>
<td>20332</td>
<td>16.21922317</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fidelia</td>
<td>19794</td>
<td>257.8196954</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LSTM</td>
<td>19794</td>
<td>257.8196954</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>mean-field</td>
<td>19794</td>
<td>257.8196954</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fidelia</td>
<td>49872</td>
<td>1.59185495</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LSTM</td>
<td>49782</td>
<td>152.3689971</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>mean-field</td>
<td>49782</td>
<td>152.3689971</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fidelia</td>
<td>80222</td>
<td>14.55081707</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LSTM</td>
<td>80184</td>
<td>152.3689971</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>mean-field</td>
<td>80184</td>
<td>152.3689971</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 4. Table 2, with standard deviations.

<table>
<thead>
<tr>
<th>Program</th>
<th>Structures</th>
<th>#RVs</th>
<th>LoC (model)</th>
<th>Capacity</th>
<th>Validation Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Mean-Field</td>
</tr>
<tr>
<td>treebn</td>
<td>BN</td>
<td>31</td>
<td>37</td>
<td>2K</td>
<td>51.88 ± 0.17</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5K</td>
<td>51.88 ± 0.17</td>
</tr>
<tr>
<td>ecoli70</td>
<td>BN</td>
<td>46</td>
<td>54</td>
<td>10K</td>
<td>11.89 ± 0.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>15K</td>
<td>11.94 ± 0.13</td>
</tr>
<tr>
<td>longrange (k = 5)</td>
<td>loop</td>
<td>8</td>
<td>16</td>
<td>1.4K</td>
<td>25.39 ± 0.00</td>
</tr>
<tr>
<td>longrange (k = 10)</td>
<td>loop</td>
<td>13</td>
<td>16</td>
<td>1.4K</td>
<td>44.05 ± 0.00</td>
</tr>
<tr>
<td>longrange (k = 20)</td>
<td>loop</td>
<td>23</td>
<td>16</td>
<td>1.4K</td>
<td>81.23 ± 0.01</td>
</tr>
<tr>
<td>longrange (k = 40)</td>
<td>loop</td>
<td>43</td>
<td>16</td>
<td>1.4K</td>
<td>155.79 ± 0.01</td>
</tr>
<tr>
<td>captcha</td>
<td>recursion</td>
<td>6</td>
<td>43</td>
<td>4.4M</td>
<td>1.644 ± 0.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>astropcfg</td>
<td>recursion</td>
<td>15.0</td>
<td>159</td>
<td>100K</td>
<td>6.23 ± 0.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>175K</td>
<td>6.22 ± 0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>250K</td>
<td>6.22 ± 0.00</td>
</tr>
<tr>
<td>mathcaptcha</td>
<td>recursion</td>
<td>14.5</td>
<td>237</td>
<td>100K+1M(CNN)</td>
<td>7.96 ± 0.46</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>250K+1M(CNN)</td>
<td>8.15 ± 0.37</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>500K+1M(CNN)</td>
<td>8.33 ± 0.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2M+1M(CNN)</td>
<td>8.39 ± 0.60</td>
</tr>
<tr>
<td>gmmcc</td>
<td>recursion</td>
<td>217.4</td>
<td>49</td>
<td>20K</td>
<td>111.11 ± 0.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>50K</td>
<td>111.16 ± 0.34</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>80K</td>
<td>111.06 ± 0.32</td>
</tr>
</tbody>
</table>

Table 5. Table in Figure 17, with standard deviations.

<table>
<thead>
<tr>
<th>k</th>
<th>ESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1.53 ± 0.69</td>
</tr>
<tr>
<td>10</td>
<td>1.00 ± 0.00</td>
</tr>
<tr>
<td>20</td>
<td>2.19 ± 0.54</td>
</tr>
<tr>
<td>40</td>
<td>2.66 ± 0.73</td>
</tr>
</tbody>
</table>

Table 6. Table in Figure 18, with standard deviations.

<table>
<thead>
<tr>
<th>Capacity</th>
<th>Effective Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M.-F.</td>
</tr>
<tr>
<td>2K</td>
<td>1.39 ± 0.56</td>
</tr>
<tr>
<td>5K</td>
<td>1.38 ± 0.57</td>
</tr>
</tbody>
</table>

Table 7. Table in Figure 19, with standard deviations.

<table>
<thead>
<tr>
<th>Capacity</th>
<th>Regeneration Rate</th>
<th>Levenshtein Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M.-F.</td>
<td>LSTM</td>
</tr>
<tr>
<td>100K</td>
<td>40.00% ± 0.45%</td>
<td>57.37% ± 2.46%</td>
</tr>
<tr>
<td>175K</td>
<td>39.83% ± 0.31%</td>
<td>60.83% ± 0.73%</td>
</tr>
<tr>
<td>250K</td>
<td>39.17% ± 0.12%</td>
<td>62.42% ± 4.16%</td>
</tr>
</tbody>
</table>
Table 8. Table in Figure 20, with standard deviations.

<table>
<thead>
<tr>
<th>Capacity</th>
<th>Recognition Rate</th>
<th>Levenshtein Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M.-F.</td>
<td>LSTM</td>
</tr>
<tr>
<td>100K+1M(CNN)</td>
<td>35.81% ± 2.66%</td>
<td>8.94% ± 0.96%</td>
</tr>
<tr>
<td>250K+1M(CNN)</td>
<td>36.75% ± 1.99%</td>
<td>9.82% ± 0.73%</td>
</tr>
<tr>
<td>500K+1M(CNN)</td>
<td>35.26% ± 3.42%</td>
<td>9.67% ± 0.59%</td>
</tr>
<tr>
<td>2M+1M(CNN)</td>
<td>34.85% ± 3.72%</td>
<td>39.35 ± 29.93%</td>
</tr>
</tbody>
</table>
Figure 28. Experimental results for ecoli70. Left: Training steps vs. training loss profile, with all guides having capacity 10K. Right: Quality of IS samples as measured by ESS.

<table>
<thead>
<tr>
<th>Capacity</th>
<th>ESS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M.-F.</td>
</tr>
<tr>
<td>10K</td>
<td>4.1 ± 2.2</td>
</tr>
<tr>
<td>15K</td>
<td>5.6 ± 3.5</td>
</tr>
</tbody>
</table>

Figure 29. Experimental results for captcha. All guides have capacity 4.4M.

<table>
<thead>
<tr>
<th>Recognition Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>M.-F.</td>
</tr>
<tr>
<td>LSTM</td>
</tr>
<tr>
<td>Fidelio</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>25.28% ± 0.54%</td>
</tr>
<tr>
<td>78.72% ± 0.90%</td>
</tr>
<tr>
<td>77.64% ± 1.08%</td>
</tr>
</tbody>
</table>